The Report committee for Andrew Lloyd Gordon Certifies that

this is the approved version of the following report:

The Fluviageny, a method for analyzing temporal river fragmentation using phylogenetics

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor: _____

James Howison

_____

David Arctur

The Fluviageny, a method for analyzing temporal river fragmentation using phylogenetics

by

Andrew Lloyd Gordon, B.A.

Report

Presented to the Faculty of the Graduate School
of the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in Information Studies

The University of Texas at Austin

May 2015

ACKNOWLEDGEMENTS

I'd like to acknowledge Dr. James Howison (advisor) for his help with coding concepts and for motivating me in my research and Dr. David Arctur (co-advisor) for his guidance and for inspiring me to become interested in geographic information systems research. I'd like to thank Dr. Dean Hendrickson for introducing me to the theory of fluviagenies, supporting me throughout this project, and helping me use my skills to bring fluviagenies from a concept into a reality. I'd like to acknowledge Adam Cohen and Ben Labay of the University of Texas Biodiversity Collections for providing me with data for my project and helping me with related presentations. I'd also like to thank Dr. Timothy Whiteaker, Research Scientist at the Center for Research in Water Resources at the University of Texas at Austin, for his advice and guidance in using ArcGIS and for helping to develop fluviageny tools and documentation for use in my research and future research.

The Fluviageny, a method for analyzing temporal river fragmentation using phylogenetics

by

Andrew Lloyd Gordon, MSINfoStds
The University of Texas at Austin, 2015
SUPERVISOR: James Howison

Phylogenetic trees have historically been used to determine evolutionary relatedness between organisms. In the past few decades, as we've developed increasingly powerful computational algorithms and toolsets for performing analyses using phylogenetic methods, the use of these trees has expanded into other areas, including biodiversity informatics and geoinformatics. This report proposes using phylogenetic methods to create "fluviagenies" - trees that represent the effects of river fragmentation over time caused by damming. Faculty at the Center for Research in Water Resources at the University of Texas worked to develop tools and documentation for automating the creation of river segment codes (a.k.a., "fluvcodes") based on spatiotemporal data. Python was used to generate fluviageny trees from lists of these codes. The resulting trees can be exported into the appropriate data format for use with various phylogenetics programs. The Fishes of Texas Database (fishesoftexas.org), a comprehensive geospatial database of Texas fish occurrences aggregated and normalized from 42 museum collections around the world, was employed to create an example of how this tool might be used to analyze and hypothesize changes in fish populations as a consequence of river fragmentation. Additionally, this paper serves to theorize and analyze past and future potential uses for phylogenetic trees in various other fields of informatics.

TABLE OF CONTENTS

INTRODUCTION

For over a century, phylogenetics has been used as a tool for allowing researchers to analyze relationships between species and their evolution over time. The "tree of life", which represents a link between all organisms on the planet and their relationships between one another, has served as a basis for the creation of smaller phylogenetic trees, which are manageable visual representations of this historical relationship between known species and can be used to hypothesize past and future speciation (Wiley & Lieberman, 2011). In recent decades, with the advent of more accessible computer technology, the use of phylogenetics has expanded well beyond the scope of its original intent. As more phylogenetic data has become available, the algorithmic power required to create and manipulate large phylogenetic trees has grown exponentially (Stamatakis, 2005). As a result, powerful computational algorithms and user-friendly packages have grown in number, allowing phylogenetic researchers to perform massive quantitative computations and analyses between larger sets of data than ever before. Additionally, those outside of specialized phylogenetics research have begun to use these tools to manipulate and analyze phylogenetic trees for unique purposes.

Phylogenetics in Informatics

Computational phylogenetics has become an important tool for research in various fields of informatics well beyond its traditional use in evolutionary analysis and bioinformatics, such as in ecological informatics, biodiversity informatics and, more recently, geoinformatics. As systems for phylogenetic analysis within these branches of informatics continue to appear at a never before seen pace, it's important for standards and universal metadata structures to exist that link various types of data with phylogenetics so that the tools will continue to be compatible with the data (Sarkar, 2007). The use of phylogenetic trees in bioinformatics has existed for longer than in most other fields of

informatics and serves as a prime example of the extension of traditional phylogenetics, i.e., analyzing species relationships, into other applications. For example, molecular phylogenetics is a branch of phylogenetics that traces relationships between sequences of DNA (Murphy, 2001). Wu, et al., developed phylogenetic algorithms for identifying large numbers of single-copy orthologous genes (COSII) for comparative and systematic studies (Wu, Mueller, Crouzillat, Petiard, & Tanksley, 2006, p. 1407).

A plethora of other free tools have been made for phylogenetic analysis in recent years. TNT, a cross-platform program with an interface that can run under Windows, allows efficient analysis of large phylogeny data sets and several other features, including methods for diagnosing and exploring trees and generating larger tree-diagrams (Goloboff, Farris, & Nixon, 2008). CDAO-Store ontologies were created to "[facilitate] the storage and retrieval of phylogenetic data" and provide "semantic descriptions of data and transformations commonly found in the domain of phylogenetic analysis" (Chisham, Wright, Le, Son, & Pontelli, 2011, p. 1). Additionally, specialized packages for R (R Development Core Team, 2008) and Python (PythonLabs, 2014) are being used for certain applications of phylogenetics within various contexts, such as bioinformatics research. The tools mentioned above have focused on being open-source and promoting the sharing of data and algorithms between researchers to facilitate the generation of quality research. One of the more difficult problems encountered in the development of phylogenetic algorithms is that of information visualization (Stamatakis, 2005). New tools are continuously being developed to solve issues of visualizing large-scale phylogenetic trees.

Biodiversity informatics researchers have become popular users of phylogenetic methods, as larger collections of biodiversity data are becoming more readily available; however, it's an emerging discipline and doesn't have a strong foundation in phylogenetics compared to bioinformatics (Sarkar, 2007). Recently, new tools have been proposed and created that bridge the gap between

biodiversity informatics and phylogenetics, such as PhyloJIVE, which integrates biodiversity and phylogenetic data for use in a graphical interface (Miller & Jolley-Rogers, 2014). With the advent of Web 2.0, biological taxa data is becoming more widely available and standardized across various fields of research; to keep up with this, new resources are appearing at an alarming rate (Penev, Roberts, Smith, Agost, & Erwin, 2010). Ecological informatics has similarly started relying on phylogenetic data to analyze species changes as a result of environmental shifts, such as climate change. Recently, as the number of online phylogenetics publications has increased compared to those in the TreeBASE phylogenetic repository, there has been a shift towards filling in the data gaps for evolutionary informatics by linking evolutionary data across the Tree of Life (Parr, Guralnick, Cellinese, & Page, 2012). New dynamic null models are also being created for addressing over-reliance on using statistical null models to apply phylogenetics to analyzing ecological community structure (Pigot & Etienne, 2015, p. 1).

Geoinformatics has only recently started incorporating the use of phylogenetics. A significant obstacle for biodiversity researchers and geoinformaticians is synthesizing and sharing their data using visual means, which is often accomplished using geographic information systems (Guralnick & Hill, 2008). However, visualization of phylogenetic trees, especially within a geographic context, is far more complex. New methods of visualization have recently come to light, such as the Global Position Trees (GPT) geophylogeny, which allow users to map phylogenetic trees on Google Earth based on species locations (Puigbo & Major, 2015). Since the late 1980s when the term "phylogeography" was first coined, a large number of theories, methods, and tools have been developed for computing and analyzing phylogeographic information. Other recent methods of visualizing this type of information include three-dimensional tree viewers, tree folding, and multi-monitor displays (Page, 2011). A unique application of phylogenetics in geoinformatics is to implement concepts of shared histories onto a geographic and temporal scale to describe

local and wide-scale changes in landscapes and ecosystems. Particularly, biodiversity informatics and ecological informatics use of phylogenetic principles can be applied on a geographic scale. For example, a dataset of historical species changes in a specific region can be analyzed using phylogenetic trees, which can be compared to the evolutionary history of the landscape itself in the form of another tree. This is particularly useful because, as Cavender-Bares, et al., state, the recent explosion in the development of phylogenetic algorithms has "revived historical traditions integrating ecology and evolution" and "increasingly demonstrates that a legacy of evolutionary history persists in ecological patterns... in concert with growing evidence for rapid adaptive evolution of populations in response to recent environmental change" (Cavender-Bares, Ackerly, & Kozak, 2012, p. 1).

Geoinformatics: Challenges with Spatiotemporal Data

One of the most challenging types of data to visualize in geoinformatics studies is spatiotemporal data, i.e., information that consists of both a spatial aspect, such as geographic coordinates, and a temporal one, such as dates. Typically, it's easier to analyze data using only one of these attributes. For example, temporally, if researchers are interested in changes in species populations over time, this can be done when focused on a fixed geographic region. Spatially, if they are interested finding patterns in species changes within various regions at a given point in time, this can easily be achieved. These types of analyses become more complex when considering both space and time, especially with changes in spatial data that occur across time, such as a landscape's evolution over time affecting the evolution of species within its various habitats. Creating appropriate algorithms to solve such problems, i.e., for processing and visualizing spatiotemporal data, first became problematic in the late 1970s. In the 1980s, temporal data handling was incorporated into database management systems but was not capable of handling the complexity of

combined spatial and temporal data. It wasn't until implementation of geographic information systems in the late 80s and early 90s that solutions to processing and presenting spatiotemporal data came to light. However, there were still severe limitations to these systems, especially when it came to representing space and time with multiple attributes simultaneously. In the early 2000s, two views of spatiotemporal data modeling were theorized: the discrete and continuous views. Discrete views represent spatial and temporal data as attributes attached to an entity (i.e., object) while continuous views denote objects as attributes attached to space-time (Peuquet, 2001).

Currently, one of the more common ways to visualize spatiotemporal data is still tied to using geographic information systems. Previously, GIS was not capable of visualizing the temporal aspect of spatial data but as new models for handling this data were proposed and developed, such as the event-based spatiotemporal data model, a.k.a., 'ESTDM' (Peuquet & Duan, 1995), the capabilities of these systems began to encompass spatiotemporal data. ArcGIS (ESRI, 2013) has historically been used to map spatial data, such as hydrological networks, and now also has tools available for visualizing temporal data using the time series, feature series, attribute series and raster series components, the latter allowing details of spatial variables to be plotted three-dimensionally along temporal variables (Goodall, Maidment, & Sorenson, 2004). However, these tools are not always effective and may be difficult to comprehend visually, especially by non-domain experts. Another GIS-based approach is the GSTP (generalized space-time path), which contains small sets of representative space-time paths derived from raw data by “identifying spatial cluster centers of observed individuals at different time periods and connecting them according to their temporal sequence” (Shaw, Yu, Bombom, 2008). Yu and Shaw also designed a GIS-based spatiotemporal visualization system based on information written in a travel diary, which contained both spatial and temporal information along with attributes. Information within the diaries was placed into a space-time path for

each individual person and integrated three-dimensionally into ArcGIS 8 (Yu and Shaw, 2004).

Various other tools exist outside of ArcGIS for visualizing and analyzing spatiotemporal data. Data mining algorithms and specialized index structures for handling spatiotemporal data have been fleshed out using periodic pattern analysis (Mamoulis, et al., 2004). R-trees, which have been used for representing spatial objects, were proposed and conceptualized to handle historical data, i.e., spatial information (Nascimento & Silva, 1998). More recently, other types of trees have been used as tools for spatiotemporal analysis, such as the Po-tree, which is based on differentiation of temporal and spatial data (Noel, Servigne, & Laurini, 2005) and the spatiotemporal relational probability trees (SRPTs), which can be used to solve complex algorithmic problems with large-scale spatiotemporal data sets, such as meteorological data (McGovern, Hiers, Collier, Gagne II, & Brown, 2008). However, these tools don't serve as effective methods for visualization. For analyzing patterns and presenting findings, it is important that such tools be capable of handling spatiotemporal data algorithmically and displaying it visually. Spatiotemporal data visualization is important in ecological-based geoinformatics studies for viewing changes to habitats and landscapes over time, including river fragmentation caused by the introduction of anthropogenic and natural barriers.

The Effects of Damming on Ecosystems

Damming for water resource management has been common worldwide over the last century (Fukushima, Kameyama, Kaneko, Nakao, & Steel, 2007). Construction of these dams fragments riverine ecosystems into segments that become isolated from one another and evolve separately (Dynesius & Nilsson, 1994). 85% of rivers within the conterminous United States are fragmented by impoundments (Perkin & Gido, 2011) and the United States has led the world in dam building for the past 100 years (Bowman, 2002). As a result, fish populations

as a whole have declined drastically over the last 50 years (Fausch, Torgersen, Baxter, & Li, 2002) and nearly 40% of North American fishes are imperiled (Jelks, et al., 2008).

In the United States, dam construction has decreased significantly over the last few decades (Fukushima, Kameyama, Kaneko, Nakao, & Steel, 2007) and dams are increasingly being removed, which reconnects previously fragmented ecosystems and aims to restore the adverse effects dams had on the structure of the river's ecosystem (Poff & Hart, 2002). Studies on the cumulative ecological effects of damming and fragmentation have received less attention than individual dam studies (Poff & Hart, 2002) and very few studies have incorporated spatial analysis across vast geographic regions (Cooper, 2013). Large dams are being researched the most while medium and small sized dams are less understood (Chin, Laurencio, & Martinez, 2011), even as dam removal has been targeting these smaller dams (Graph, 2005). Large dam removal is less common and thus it is expected that their impacts via fragmentation are more pronounced and will continue farther into the future. While large dams have a significant impact on local ecosystems, small and medium sized dams comprise 97% of all dams in the state of Texas (Chin, Laurencio, & Martinez, 2011), which is heavily dammed (Figure A1). Dams on large, central rivers within watersheds tend to exert the greatest influence on fragmentation (Cooper, 2013).

Downstream effects of dams have received the most attention from researchers but upstream impacts can be just as significant (Greathouse, Pringle, McDowell, & Holmquist, 2006). For example, two imperiled species, Notropis oxyrhynchus and Notropis buccula of the Brazos River Basin in Texas, have all but completely disappeared from the upper majority of the basin (Wilde & Urbancyzk, 2013). Downstream disturbances can transmit to far upstream reaches and affect reproduction and patterns of biodiversity (Pringle, 1997). In some cases, historical effects of damming on fish populations may not be apparent on regional levels but may only be evident on smaller scales, e.g., as

shown in recent data from Oklahoma and Western Arkansas (Matthews & Marsh-Matthews, 2015).

Long-term temporal analyses of damming and resulting changes in fish populations are important for making management decisions; the efficacy of short-term analysis is questionable compared to long-term analysis (Ligon, et al., 1995). Fisheries management activities are usually narrow in scope and only focus on a specific area or short stretch of river (Zorn, Seelbach, & Wiley, 2002), so understanding damming trends along entire river systems allows current projects, such as removal and restoration, to be assessed and prioritized on a larger scale and may influence future management decisions (Hall, Jordaan, & Frisk, 2011). Learning from past experiences with dam construction can guide future decisions on dam removal (Babbitt, 2002). With the introduction of richer and more accurate hydrological data, such as the National Anthropogenic Barriers Dataset (Ostroff, Wieferich, Cooper, & Infante, 2013) and the Global Water Bodies database of high-resolution lake imagery (Verpoorter, Kutser, Seekell, & Tranvik, 2014) along with larger, standardized species data, such as the Fishes of Texas database (Hendrickson & Cohen, 2010), these studies on the effects of damming are becoming more reliable.

Natural History Collections as Tools for Research

In order to conduct research on how changes in environment, such as damming, affect habitats and species, a large amount of reliable data must be used. In most cases, it's difficult to obtain such rich data, especially in small-scale, local studies. One of the most important resources for such data is the natural history museum. Natural history collections have been electronically cataloged since the 1970s but computerized specimen data still only accounts for a small fraction of natural history museum collections (Graham, Ferrier, Heuttman, Moritz, & Peterson, 2004). Much of this lack of digitization can be attributed to a lack of funding for natural history museums; as a result, irregular electronic

availability of collections data leads to a significant drop in this data's usefulness for research (Vollmar, Macklin, & Ford, 2010). Field notes from collection instances are an example of valuable pieces of data that are widely left out of the digitization process or may only be summarized in cataloged descriptions. High-resolution pictures of cataloged specimens can also be important for giving researchers in any geographic location easy access to information regarding each collection object. It's vital for natural history museums to digitize as much of this information as possible, as they are a valuable resources for several fields of research, including ecology, evolutionary biology and conservation (Graham, Ferrier, Heuttman, Moritz, & Peterson, 2004).

There are other significant issues with using natural history collection data besides the lack of electronic availability, such as "taxonomic inaccuracies and biases in the spatial coverage of data" and unreliability in certain types of data such as absence of species at locations; this absence doesn't necessarily imply true geographic absence of species, it only means that the collectors didn't find the species there at that given time (Graham, Ferrier, Heuttman, Moritz, & Peterson, 2004, p.497). However, over the years, the quality of natural history collection data has improved in certain data integrity projects, such as the Fishes of Texas Database (Hendrickson & Cohen, 2010), which synthesizes Texas fish collection data from institutions worldwide and aims to fix taxonomic errors and fill geographic gaps in data. The Global Biodiversity Information Facility (GBIF) and the Taxonomic Databases Working Group (TDWG) for Biodiversity Information Standards have been working towards improving poor quality of data in many natural history collections from hundreds of sources, including generation of standardized metadata vocabularies and online integration to include suggested improvements by user communities (Lapp, Morris, Catapano, Hober, & Morrison, 2011). Ponder, et al., used statistical calculations for describing separation, density, and clustering of sampling points, alongside geographic information systems, to model predictions in coverage of certain

species in order to improve gaps in data for species distributions (Ponder, Carter, Flemons, & Chapman, 2010).

Despite its shortfalls, museum collection data is still vital to certain types of specimen-based research, especially geographic-based analyses such as species richness and species diversity calculations along geographic gradients, assuming that there are not significant gaps in sampling data (Grytness & Romdal, 2008). Large-scale analyses of species changes, such as declines of species over time, require and have already been successful with the use of historical natural history collections data, since new data from the past cannot be collected due to the temporal factor (Shaffer, Fisher, & Davidson, 1998). Such research will further benefit from higher quality collections data as museums push for integrating their data into larger databases, normalizing, and improving the accuracy and reliability of this data. Museum collections continue to be the most comprehensive source of information available for biodiversity research, even with the presence of their significant flaws, such as gaps in digitization and geographic coverage (Ponder, Carter, Flemons, & Chapman, 2010).

Data sharing is also crucial to improving the quantity of available data and its accessibility. Promoting open data is an effective way of aiding scientific publishing; there are many benefits to repositories that are shared amongst the scientific community, such as Dryad, which allows authors to archive their data and encourages other researchers to use this data (Vision, 2010). In bioinformatics, commitments to responsible sharing of academic research health data between academic institutions is promoted and found to be effective in facilitating research (Piwowar, et al. 2008). It's important that data aggregated from multiple sources is normalized and maintained to improve accuracy and consistency for searching and research purposes. For example, differences in cataloging practices between universities or museums may lead to inconsistencies in or absence of appropriate metadata. While some institutions may use a fine granularity of metadata for their specimens, such as field notes,

photographs, or the gear used to collect them, others may not include this information. There may also be differences in metadata terms such as abbreviations for terminologies or rules and guidelines for naming and georeferencing collection object locations. VertNet (www.vertnet.org), a cloud-based system for sharing biodiversity data, promotes open access and improvement of shared scientific data by monitoring data contributions, performing quality assessments on these contributions, standardizing them using controlled vocabularies, and georeferencing location data. It also allows consumers of the service to make annotations to data to improve its quality and reliability (Constable, et al., 2010). Even social media tools such as Facebook have been found to aid researchers in improving accuracy of data as a result of other researchers helping to identifying species of fish from photographs posted publicly on the site (Sidlauskas, et al., 2011).

Fluviagenies

With the increase in availability of sufficient quantity of reliable species data, coupled with ample geographic spatial and temporal data, research on damming impacts can now benefit from more appropriate tools for conducting analyses on this data. Only recently have studies on ecological impacts of damming begun to incorporate a temporal measure alongside a spatial one, typically using geographic information systems to accomplish this. The effects of dams on river ecosystems is so complex that there is a need to account for both temporal and spatial factors when conducting analysis on fluvial habitat changes (Cooper, 2013). The technology required to perform temporal studies is increasing (Mann, 2012) but there are few methods for efficiently assessing temporal factors alongside spatial ones.

Robert Ewers, et al., recently conducted research using phylogenetic trees to visualize landscape fragmentation and study resulting ecosystem changes (Ewers, et al., 2013). The authors defined this method as a 'terrageny', which

quantifies the history of a landscape and shows ancestry of present day fragments by tracing from the tips of the tree back to the root source. The authors used this terrageny to compare quantified patterns of shared species between landscape fragments and to predict spatial patterns of habitat loss. Additionally, they suggested that these trees could be used to further hypothesize future species changes based on comparing patterns between terragenies with species phylogenies.

The fluviageny is a new method that we have developed based on the concept of the terrageny introduced by Ewers, et al.; it would similarly draw phylogenetic trees based on the fragmentation of river systems instead of landscapes. While these trees could be drawn as a result of any fragmentation where ample data is available, such as drought, fluviagenies could be especially useful in analyzing significant and constant fragmentation caused by damming. Fluviagenies are created in a different manner than terragenies due to the nature of riverine systems. While landscape fragmentation occurs in a geometric fashion, splitting landscapes into polygonal fragments, river fragmentation occurs with an interrupt to the downstream flow of the river or stream and subdivides it into smaller drainage basins. Thus, with each dam that is introduced into a river, two fragments of the river are created – one upstream of the barrier, and one downstream of the barrier. It's impossible to visualize the relationships between fragments created this way using only maps, because fragments may be bound by multiple dams and may exist as combinations of their ancestor fragments at different points in time (e.g., back in time before any damming existed, all fragments existed as one uninterrupted system).

Figures 1 through 4 (below) show how a fluviageny would be drawn from the construction of the first few dams in a river basin. When a river basin has no damming, and hypothetically no fragmentation at all, it is represented by a straight line across time (Figure 1). The line expresses the continuity of the basin's streamflow, which is assumed to be uninterrupted. This line can represent an uninterrupted habitat, although this is only considering whatever factors we are analyzing within the fluviageny, which in this case is only dams and does not account for other variables such as shifts in climate.

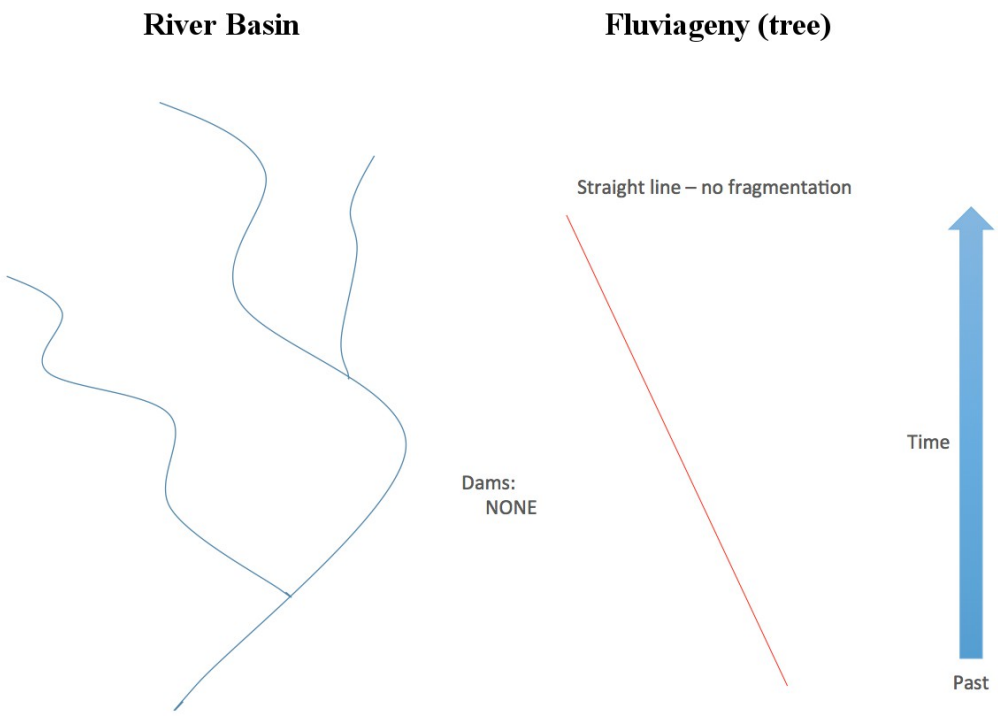

Figure 1. Left – a hypothetical drainage basin with blue lines representing rivers draining from upstream (top) to downstream (bottom). Right – a 'fluviageny' tree, representing the basin throughout time as if no fragmentation existed. The entire basin is treated as one continuous system (i.e., one line).

Once a dam is built somewhere along a river within the basin, it fragments the basin into two subbasins, one upstream of the new dam and another downstream of the new dam. This is represented on the temporal line by a binary split within the tree. The split occurs at the point (date) where the dam is constructed as a node. The node's children represent the two fragments created by the node – one upstream and the other downstream (Figure 2). There is no spatial representation within a fluviageny tree, only a spatial relationship between the dams (as nodes) and the fragments they create (as children).

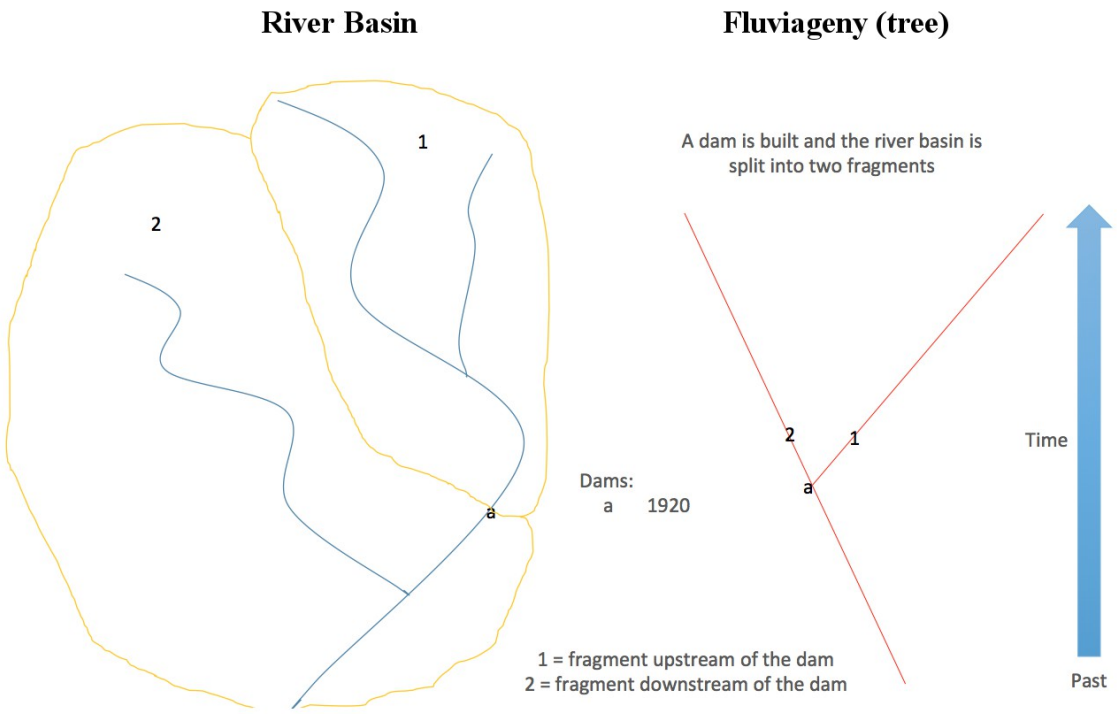

Figure 2. Left – the basin with a single dam constructed in 1920 and labeled as "a". The dam splits the basin into two fragments – one upstream (labeled "1") of the dam and another downstream (labeled "2") of the dam. Right – the fluviageny corresponding to the introduction of this dam. Now, as we progress up the fluviageny through time, when we reach 1920, the system splits from one continuous system into two. The node of the tree represents the dam ("a") and the two branches show the new fragments created as a result of the introduction of dam "a", one upstream ("1") and the other downstream ("2").

For each additional dam built sequentially in the river basin, another split occurs on one of the branches of the tree corresponding to the new dam's position with relation to the other dams and their segments (Figure 3). If a new dam occurs upstream of another dam, it will be placed somewhere on the branch corresponding to this upstream location (i.e., the branch 'upstream' of the node). Each branch serves as a linear representation of time across a spatial area. When a single dam goes into a basin it splits the basin roughly into two spatial areas – the area that drains upstream and the area that drains downstream of the dam.

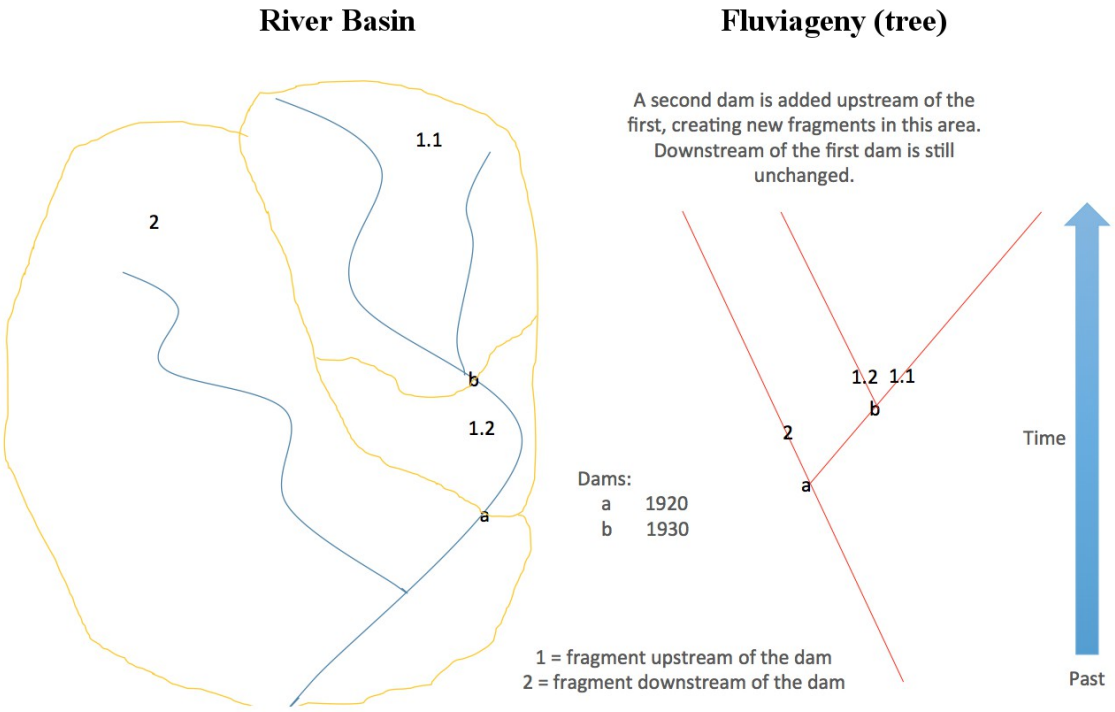

Figure 3. Left – a second dam ("b") is added to the drainage basin. The fragment upstream of dam "a" (fragment "1") is now split into two subfragments, "1.1" - upstream of dam "b", and "1.2" - downstream of dam "b". Right – dam "b" is now added to the fluviageny tree. Since dam "a" was built in 1920 and dam "b" was built in 1930, after dam "a", dam "b" occurs next in sequence, temporally, after dam "a". It is placed upstream of dam "a" on the tree (on branch "1") and creates two new branches, "1.2" and "1.1", which represent upstream and downstream of dam "b".

When a fluviageny tree has been completely drawn, each tip of the tree represents the finest level of fragmentation based on the data used (Figure 4). If there are five dams used in the data to create a fluviageny, six fragments should be created, i.e., there should be six leaves and thus six tips to the tree. The ends of these tips represents the 'cut-off date' for dams being built, which would likely be the present day but could be modified based on research needs.

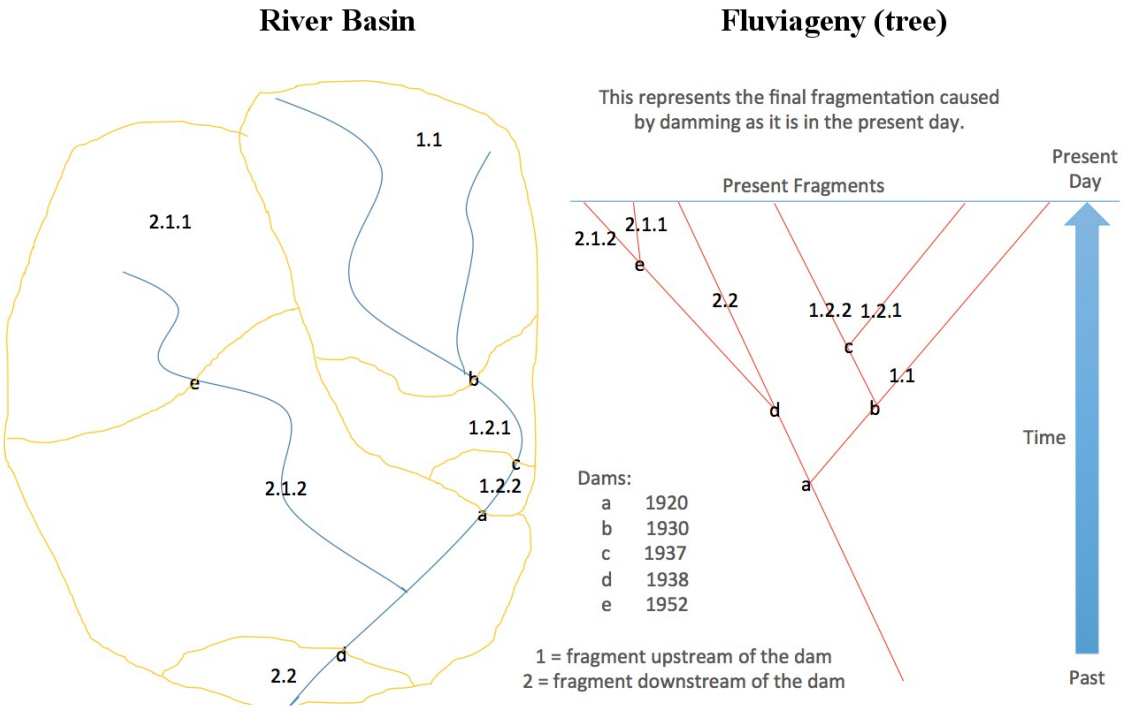

Figure 4. Left – the drainage basin fragmented by five dams. Each fragment represents a portion of the river basin and is labeled according to its upstream/downstream relationship to the dams that directly impact that fragment, where "1" is upstream and "2" is downstream. Right - A fluviageny of a hypothetical present day basin, containing five total dams. The tips of all branches in the fluviageny represent the present day fragments created as a result of the five dams constructed along the river network. The nodes represent dams, which cause the tree to split each time into "upstream" and "downstream" subfragments of the previous ancestral fragment. Dams are placed along the tree in order, temporally, from root to tip, according to when they were constructed. All fragments can be traced back to one common ancestor – an uninterrupted system with no splits, representing the entire basin before any fragmentation occurred.

In order to build fluviagenies, fragments created by the dams in question must be labeled using fluviageny codes or "fluvcodes". Fluvcodes consist of a sequence of 1's and 2's representing the relationship between fragments and dams. Each 1 represents an upstream relationship with a dam and each 2 represents a downstream relationship. As each dam is added to a basin, a new split occurs and 1's and 2's are added to the fragments (in the case of the first dam built) or appended on to the previous 1's and 2's based on the location with respect to the previous dams constructed. Figures 1 through 4 contain these fluvcodes, shown both in the tree and on the river basin.

If more than one dam is built in the same time frame, i.e., if the finest level of temporal metadata we have is the year it was built and two dams were built in the same year, then we produce an unresolved polytomy at the node representing the dam on the tree (not shown in the above figures). The polytomy, instead of creating a binary split at the node into two children, would create an additional child for each additional dam, producing three branches per node for two dams in the same date, four branches for three dams, etc. With studies that are limited in scale, either geographically or temporally, or those limited to a specific selection of barriers, such as large-classified dams, it is less likely to encounter these unresolved polytomies. However, in larger-scale fluviagenies, these would be fairly common. They must be resolved manually by obtaining finer temporal data for each dam and appropriately adjusting the scale of the resulting fluviagenies. For research purposes, this would likely not have any significant impacts on using fluviagenies for performing quantitative analysis, as dams built within very short periods of time would likely have little impact within those slim time slots.

While most spatiotemporal data is seen in three-dimensions, fluviagenies allow easy-to-understand visualization of certain types of spatiotemporal data in only two dimensions. A fluviageny could even be used to allow visualization of non-spatial data that contains two attributes: a temporal aspect, and a split in

data with relationship to past ancestral fragments in the data. It is also easy to visualize the relationship between fragments within a fluviageny, as the leaves of the tree can be traced back, connecting fragments of a basin with their ancestral fragments, and thus tracing the history of the river system.

As mentioned previously, spatiotemporal data has previously been categorized as either being in the discrete or continuous view (Peuquet, 2001). Fluviagenies transform this discrete view into a continuous view, taking entities (dams and rivers) with spatial and temporal attributes (points, locations, and dates) and transforming them into associations based on spatial relationships across a temporal 'map', represented by a phylogenetic tree. The tree starts at any given point in time and ends at any given point in time. Objects (dams) exist as nodes on the tree and form branches based on spatial relationships and as one traces across the tree, they move through time. While the tree does not provide complete visual detail of the spatial existence of this data, as would be seen on a map, it does show the spatial relationship between these data.

METHODS

Data and Software

This study used four sets of data: base maps, river shapefiles, dam shapefiles, and fish information in the form of spreadsheets, from four separate sources. The base maps were downloaded from the Texas Parks and Wildlife Department's GIS Lab Data Downloads (TPWD, 2012). This includes the maps for Texas and its counties and basins. The shapefiles for the rivers come from hydro maps downloaded from the Texas Commission on Environmental Quality's website (TCEQ, 2013). The hydro maps represent only perennial water (year-round) within Texas. While some TCEQ dam data was used to show dam coverage in the state of Texas, the remaining dam data was downloaded from the National Anthropogenic Barrier Database, or NABD (Ostroff, Wieferich, Cooper, & Infante, 2013), which is linked to the National Hydrography Dataset Plus, Version 1 (NHDPlusV1) and National Inventory of Dams (NID) datasets. The NABD fills in large gaps that exist within the NID, such as a lack of small dams (Poff and Hart, 2002) and an overall underrepresentation of dams (Chin, Laurencio, & Martinez, 2011). NABD dam data was used in conjunction with TCEQ data for fluviageny analysis and tree creation. Fish collecting instance information was retrieved from the University of Texas Biodiversity Collection's Track 2 dataset (University of Texas, 2015) and the Fishes of Texas database (Hendrickson & Cohen, 2010) as Excel spreadsheets. NABD and fish data were further cleaned in Excel by correcting typos, removing exact duplicates, and replacing old species identification types with newer, more accurate information.

Much of the fish data used in this study is a part of the Fishes of Texas Project, which is a database hosted by the University of Texas Biodiversity Collections (formerly Texas Natural History Collections) and a website hosted by the Texas Advanced Computing Center. It is designed to combine fish data from across the state of Texas and donors outside of the state to provide researchers with normalized, reliable data (Hendrickson & Cohen, 2010). It contains records

from over 40 institutions dating back to the 1800s. All data is georeferenced and standardized, with error detection and corrections made to ensure accuracy. Using this large dataset is vital to temporal and spatial analysis, which requires a sufficient sample size not only across different regions but also across different time periods. Before the creation of this database, it was difficult to conduct general, large-scale analysis on fish assemblages in Texas due to the potential for inaccurate information that may be difficult to access due to disparate sources and format inconsistencies (Hendrickson & Cohen, 2010). The project continues into the future with database improvement, addition of new data, and filling in of data gaps. Unfortunately, for temporal analyses, filling data gaps can be difficult since we cannot go back in time to collect fish. However, older, uncataloged specimens exist that could provide valuable information, once recorded.

ArcMap 10.1, part of ArcGIS, (ESRI, 2013) was employed for the creation of all maps. Dr. Timothy Whiteaker's fluviageny toolkit and documentation were used within ArcGIS for generation of fluviageny codes as assigned to river segments between dams (Whiteaker, 2014), which also requires the Arc Hydro Tools (Maidment, 2003). The fluvcode toolkit for use within ArcGIS, including documentation with instructions for use, can be downloaded from the University of Texas Center for Research in Water Resources tools site under “fluviageny”. The ETE2 library (Huerta-Cepas, Dopazo, & Gabaldon, 2010) in Python (PythonLabs, 2014) was used for scripting the generation of fluviageny trees. R (R Development Core Team, 2008) and the APE package (Paradis, Claude, & Strimmer, 2004) were used to create code for manual creation of miniature, time-scaled fluviagenies using adjacency matrices.

Generating Fluvcodes

The first step in creating a fluviageny from geographic and temporal data is to generate "fluvcodes". In order to create fluvcodes, appropriate data must first be retrieved and imported into ArcGIS. The minimum required data consists of (1) shapefiles for the river basin(s) being researched as a line feature class and (2) information for dams along the river system, including geographic coordinates as a point feature class. Some dam data already includes the next dam downstream of the current dam but this can easily be calculated in ArcGIS if it is not available in the dataset that is used. The data must be cleaned appropriately in ArcGIS so that there are no disconnected river segments and all dams used in the analysis intersect the main river network. Rivers must also exist as segments by ensuring that dams are located at the endpoints of rivers, which can be checked and fixed manually in ArcGIS.

Once all of the data is in place, unique identifiers are added for each dam and a field is added for the next downstream dam with respect to each dam if it does not already exist in the data (except for the furthest downstream dam or 'root' dam), which is then calculated while building a geometric network. A geometric network must be built and flow directions for all lines are set to run downstream to the end of the river basin, which can be calculated in Arc Hydro or using ArcMap's Utility Network Analyst tool. Since geometric networks require points at all junctions between river segments, these will be assigned as a new point feature class and must also be considered when labeled with fluvcodes; however, a junction is not considered a barrier and thus several river segments forming a junction may still exist in the same fluviageny fragment. The next downstream dam id, is then calculated using Arc Hydro attribute tools. Arc Hydro computes this using network tracing. The last attribute that is needed for fluvcode generation is the junction identifier. The junction identifier is assigned to the stream segments that exist in between dams and represents the next downstream dam from that stream segment. Tools for computing the junction

identifier are included in Whiteaker's fluviageny toolset. The fluviageny toolbox uses ArcMap's Model Builder to generate fluvcodes. The model traces the relationship between dams and river segments by using the primary identifiers assigned to each column and placing them into temporary dam identifier columns, which are reset at each run through the model. For more detailed instructions on generating fluvcodes, see Timothy Whiteaker's fluvcode documentation and toolbox (Whiteaker, 2014).

Building Fluviagenies

In order to build fluviagenies from fluviageny codes, the codes must be exported and run through a script in either R or Python. A Python script to generate fluviagenies from fluvcodes and corresponding dates has been written for this report; the data can be exported from ArcGIS in the form of a .csv file and run through the script. The data in the csv must consist of two columns of information: (1) the date each dam was built and (2) the fluvcode assigned to the next fragment upstream of the corresponding dam.

There are two aspects to creating fluviageny trees using the provided data. The first is to generate the tree using the fluviageny codes. This can be done without incorporating dam data, since each fluviageny code provides instructions for drawing the fragments of the tree. For example, the longest fluviageny code represents the depth of the tree. The number of characters in a fluvcode, ignoring the decimal points, represents the depth of that code, or at least the depth of the children associated with the code. If a basin only has a single dam, it will generate two fluvcodes: 1 and 2. The tree will only have 1 internal node (1 dam) and two child leaf nodes. Each child leaf node only has a depth of one, which corresponds to the number of characters in their fluvcodes. Similarly, a fluvcode of 1.2.2.1 will have a depth of 4 and if it is the longest fluvcode (or tied for longest), then the entire fluviageny will only have a depth of four. Each number also represents the creation of a child node for the previous

series of numbers. 1.2 can be traced on the tree with one upstream split at the first node and another downstream split at the next node upstream of the first node. If our tree is built from bottom (root) up and upstream (1) is left while downstream (2) is right, then 1.2.2.1 would be left, right, right, left and would include four internal nodes (i.e., dams).

When the tree is drawn algorithmically a breadth-first or depth-first approach can be used. The differences between these methods can be seen in Figure A2. In breadth-first, the program reads through the first number in all of the fluvcodes. If there is a 1 and a 2, two child nodes are created. Moving on to the next series of nodes (i.e., second number of the fluvcodes), for each previous node, more children are created for each 1 and 2 corresponding to the previous node. For each iteration through the list of fluvcodes, the algorithm must look at each fluvcode and create one child node for a 1 or 2 or an additional child node for another 1 or 2. When a leaf node is reached (i.e., the end of the fluvcode), the fluvcode can be removed from the list. This algorithm requires multiple runs through the list in order to execute. However, in a depth-first method, the code iterates through the list only once, reading each fluvcode one at a time and drawing each branch corresponding to that fluvcode to the leaf node. Since the ETE2 package handles tree browsing efficiently and the algorithm locates nodes by looking up node names, it's faster to use the depth-first method, since it only has to read through the list of fluvcodes once. Thus, the depth-first method was used in the Python code for fluviageny creation. After the tree is drawn, fluviagenies must be scaled temporally based on the dates of fragmentation, which correspond to dam construction dates. This is where the other column, dam dates, is used.

The code for drawing fluviagenies from fluvcode lists is shown in Figure A3, at the end of this section. The code is split into five primary parts. First, the csv file that was exported from ArcGIS, which must contain fluvcodes and the next downstream dam of each fluvcode, is imported into Python via a user-

defined file path. A list is created that contains these values, with each item in the list containing a pair of items: the date (dam) followed by the upstream fluvcode.

Second, using the ETE2 toolkit (Huerta-Cepas, Dopazo, & Gabaldon, 2010) an empty tree is created. The list of fluvcodes is sorted by dates, if it isn't already, and the first (earliest) date is placed as the root node (first dam constructed) of the tree. The first two children - "1" (upstream) and "2" (downstream) are created, which always exist in a fluviageny, since there will always be at least 1 dam. Before drawing the rest of the tree, the list must be re-sorted by fluvcodes, instead of dates, so that the tips draw in the proper order. Then, for each item in the list, a key (n) is created for slicing the fluvcode apart from the date for the first tree generation. N starts at the minimum number of characters that are needed in the string of a date plus a second-depth fluvcode (e.g., "19851.2"). This requires that the csv is formatted using comma delimiters, otherwise slight adjustments to iterators would need to be applied; it also assumes that dam dates only consist of four characters, i.e., a year, otherwise the csv would need to be split into two full lists for processing variant string lengths. For each of these strings, while the total characters of the string are not exceeded by n, if the currently selected string from after the date to the current n value is not already a node in the tree (e.g., if n=7, and the string is 19451.2.2.1, the value 4:7 would be 1.2), then that parent node, which is the previous 1 or 2 in the string (e.g., for 1.2, it would be 1) is given a new child of the current string. This continues, appending more characters onto the string until the end. For example, 1.2.2.1.2 would start as 1.2. If 1.2 is not in the tree (to prevent repeating it), it is added to the tree as a child of 1. Next, if 1.2.2 is not a node in the tree, it is added as a child of node 1.2. Next, if 1.2.2.1 is not in the tree, it is added as a child of 1.2.2. This continues through all fluvcodes, drawing the full fluviageny.

Third, another iteration through the list is necessary to replace fluvcodes of internal nodes with corresponding dam dates in order to scale the tree temporally. Dates must be added after the nodes have been created using

fluvcodes, otherwise the tree won't be drawn correctly. Dam dates cannot replace node names in the first loop through the list because this would prevent the children node from being able to find their parents while the tree is being constructed. During this run through the list, a key (z) is created for giving the maximum possible index for a given list item's character length, in this case, length – 1. The index is then continuously decreased as the characters are analyzed. At the first instance where a 1 is encountered, this means that this node is the first node that is downstream of the fluvcode, i.e., the dam is the first downstream dam from the fragment, so it is the dam appropriately paired with the fragment in the list. Thus, the fluvcode parent node (e.g., for 1.2.1, the parent is 1.2) is assigned the dam date value, which is only the first four letters of the string. This continue for all items in the list, replacing all internal nodes with dam dates. Again, this assumes dates will all be four letters, which should be the case unless we are dealing with species from before the year 1,000 or values that include partial dates or months, but this is not common in currently available dam data.

Fourth, the tree must be scaled temporally based on the dates of the internal nodes. The leaf nodes can be set at present day, if desired, which in this case is 2015. The tree is traversed using level-order traversal (the default in ETE), which is a breadth-first traversal. For each node encountered during traversal, if it is a leaf node, then the distance to its parent node is equal to the present year (e.g., 2015) minus the parent node name, which is the dam date. Otherwise, if it's not a leaf node (i.e., it's an internal dam node) then the distance to its parent is its name (year) minus its parent's name (year).

Finally, the tree can be written into Newick format for use with other tree generation programs. ETE2 also has a 'show' function that can be used to display and manipulate the tree in a tree viewing graphical user interface, called ETE browser, which is capable of browsing data in the tree or exporting it into a readable Newick file (Huerta-Cepas, Dopazo, & Gabaldon, 2010).

Geographic Information Systems

Analysis of data was performed in geographic information systems to show how fluviagenies were built and the potential uses for fluviagenies in conjunction with ample species collection data. Three river basins were selected for testing the creation of fluviagenies: Brazos, Colorado, and Trinity (Figure A4). These three basins were selected because they are the most heavily dammed in the state of Texas (Figure A5) and have an ample quantity of available dam data. Within each basin, dams were clipped to the river network within a 500-meter margin-of-error radius. All dams that were off of the primary, uninterrupted, perennial river network were excluded from the study. The three basins had a final dam count of 67 for the Brazos (Figure A6), 60 for the Colorado (Figure A7), and 102 for the Trinity (Figure A8). Fluvcodes were generated for each of these basins and the selected dams using Tim Whiteaker's fluviageny tools for ArcGIS (Whiteaker, 2014). The resulting tables for dams and rivers were exported from ArcGIS to spreadsheets. The dams spreadsheet contained the HydroID, which is a unique identifier for each dam in the dataset, and year complete for each dam. The rivers spreadsheet contained the JunctionID (i.e., the HydroID of the dam that is directly downstream of the river fragment) and the fluvcode for each segment. The two spreadsheets were combined so that each dam completion year was tied to a fluvcode (i.e., HydroID and JunctionID were matched, merged into the dam's date, and combined with fluvcodes). The resulting spreadsheet with two columns: dam year and fluvcode of the immediate upstream fragment, was exported to csv and run through the ETE2 package (Huerta-Cepas, Dopazo, & Gabaldon, 2010) in a Python (PythonLabs, 2014) script to create fluviagenies. The tree was analyzed using ETE2's tree browser and exported into a readable Newick format, which was run through the APE package (Paradis, Claude, & Strimmer, 2004) in R (R Development Core Team, 2008) to create more readable trees for the results analysis.

Figure A9 shows a generalization of the flow direction for the Colorado River Basin, one of three river basins selected for fluviageny creation. A line is drawn between selected dams to signify that the entire geographic area that is fragmented by the two dams, i.e., a subbasin of the larger river basin, has a flow direction from the upstream dam towards the downstream dam. For calculation of fluviageny codes in ArcGIS, if the river network topology is unsatisfactory or incomplete (e.g., fragmented), these lines can be carefully drawn as a new line feature class, making sure to avoid connecting known disjointed fragments, drawing the ends of the lines at dam points, and ensuring that the lines are drawn in the correct direction, from upstream to downstream.

Cyprinid fish collecting instances in the Colorado River Basin (Figure A10) were used to create a small example of how fish population changes can be mapped to fluviagenies , using only the 10 largest dams in order to sufficiently display the methods of performing these types of analyses. The Cyprinid family was selected for the quality and quantity of available data. The Colorado River Basin was selected for this example due to the availability of both a large, heavily dammed, continuous perennial river network, and due to the selection of Cyprinid fish data and its distribution in the Colorado River Basin.

RESULTS

Fluviageny Trees

Figure 5 shows the fluviageny tree for the Brazos River Basin. The tree is scaled horizontally by the year the dams were constructed. Several major patterns can be immediately distinguished by looking at the tree. The degree of fragmentation at any date range can be obtained by cutting out vertical segments of the tree, as seen in the orange circle drawn over the fluviageny. A staircase effect can be observed at several points in time where frequent, downstream fragmentation occurred, as seen in the red circles. Overall, there's a strong downstream fragmentation trend across the entire basin, as seen from the tree's tendency to skew downstream (up). It's important to note that the tree is only scaled temporally (horizontally) and no vertical scaling exists. Therefore, the vertically elongated staircase is no different than the vertically shorter ones. There is an apparent cessation of frequent fragmentation in the latter third of the tree, as indicated by the green box. Longer tree branches represent fragments that remain untouched for an extended period of time, i.e., the oldest fragments, based on the dams used in the analysis, as seen in the light blue boxes. Aged fragments could be used as null cases (i.e., less-fragmented) for comparison to heavily fragmented river segments. Additionally, tracing back through the tree can give us the degree of fragmentation for the node, which can also be seen in the length of the fluviageny code. A formula for this can be written as:

N = [(fluvcode string length) + 1] / 2

Where N is the degree of fragmentation. For example, the shortest fluviageny codes were only fragmented three times (1.1.1, 2.1.1, 2.2.1). The longest codes have the highest degree of fragmentation. This pattern makes it easy to visualize which parts of the basin have been affected by the most damming, which cannot be seen simply by looking at a map of dams and river networks. N can also be drawn from tree analysis applications as the depth of the node in the tree.

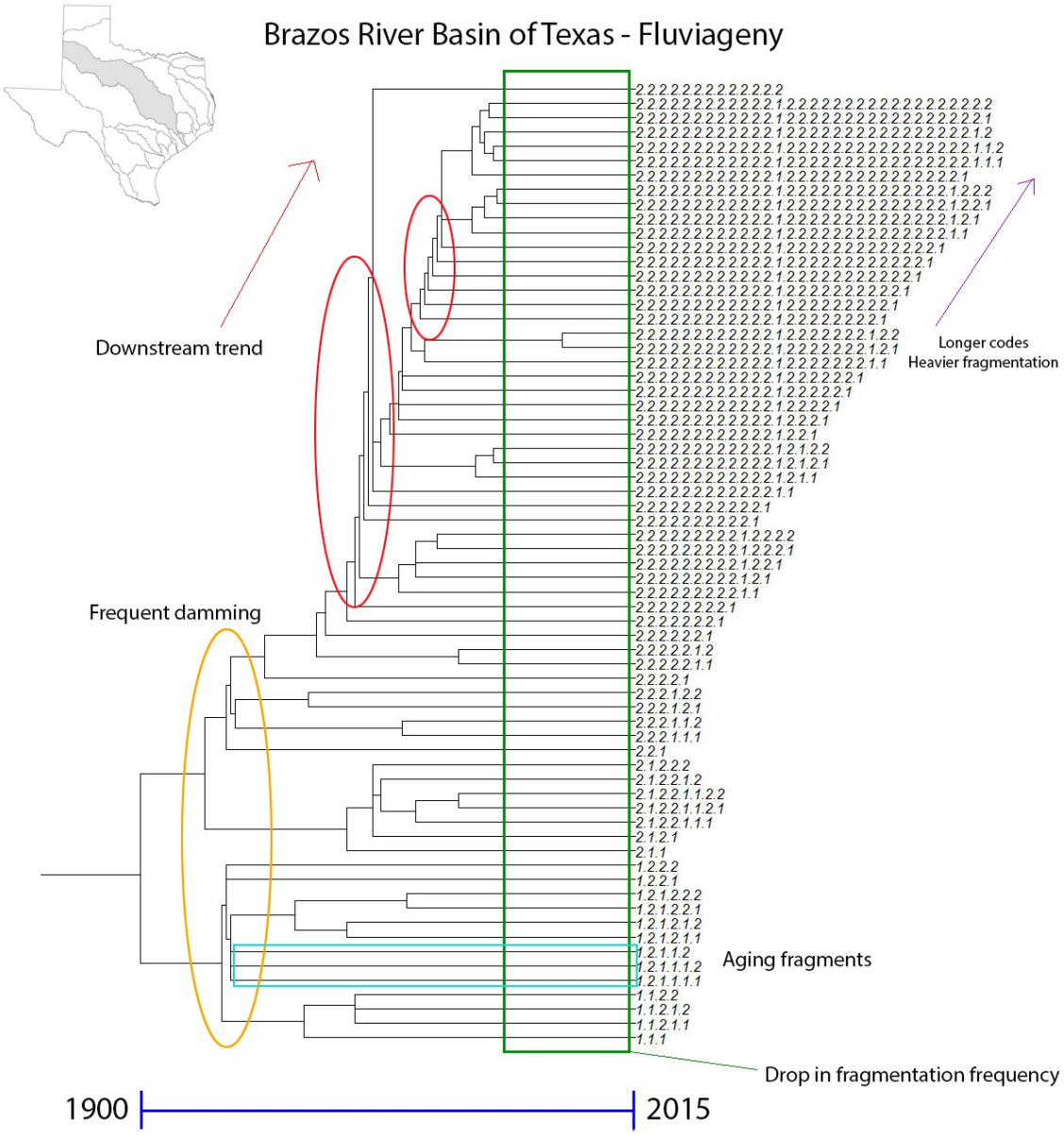

Figure 5. Fluviageny for the Brazos River Basin, using only dams along the uninterrupted primary river network.

Figure 6 shows the fluviageny tree for the Colorado River Basin. The tree is displayed similarly to the Brazos River Fluviageny, with horizontal temporal scaling and polytomies at same-date fragmentation points. The same staircase effects can be seen in this tree, as well as a higher number of old fragments and a similar cessation of frequent fragmentation. However, these patterns occur more sporadically across a wider range of dates from the early to mid 1900s.

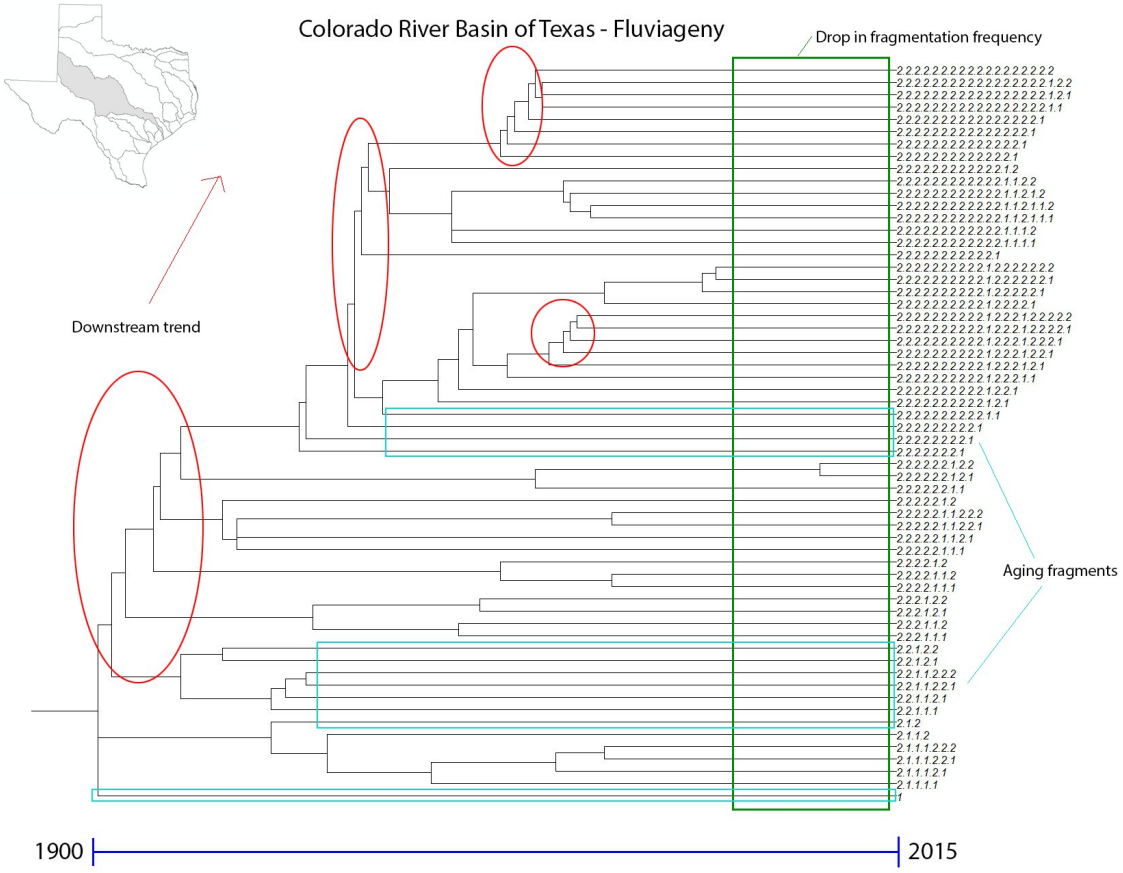

Figure 6. Fluviageny for the Colorado River Basin, using only dams along the uninterrupted primary river network.

Figure 7 shows the fluviageny tree for the Trinity River Basin. The tree is displayed similarly to the Brazos and Colorado River fluviagenies, with horizontal temporal scaling and polytomies at same-date fragmentation points. The Trinity River Basin experienced similar fragmentation patterns to the previous two basins. It has a larger number of older, less fragmented areas, as seen in the blue boxes. Interestingly enough, though not surprising based on historical damming patterns and trends, there are many cases where series of dams were placed downstream of one another in quick succession but no cases of this occurring upstream of dams. While this does not tell us anything new about damming practices, it is powerful to be able to easily see where and when this happened, which may be able to help us predict how these patterns have impacted subsequent fragmented ecosystems. Unlike the Colorado River Basin, all frequent downstream damming patterns in the Trinity River Basin fluviageny occur in a short time window around the late 1950s into the 1960s, which coincides with known damming patterns for the basin and throughout the state as well as historical damming trends in North America.

Fluviagenies can be created without polytomies for same-year dam construction dates by removing time-scaling to discern exact, non-scaled damming patterns. Figure A11 shows an example of the Colorado River Basin fluviageny without time-scaling of the branch lengths (i.e., node distances). This type of tree can be useful for visualizing the exact upstream/downstream relationship between individual dams and how fluviagenies with resolved polytomies might look. An unscaled tree with same-date nodes has these nodes compressed into polytomies, such as the first two nodes in the Colorado River Basin fluviageny.

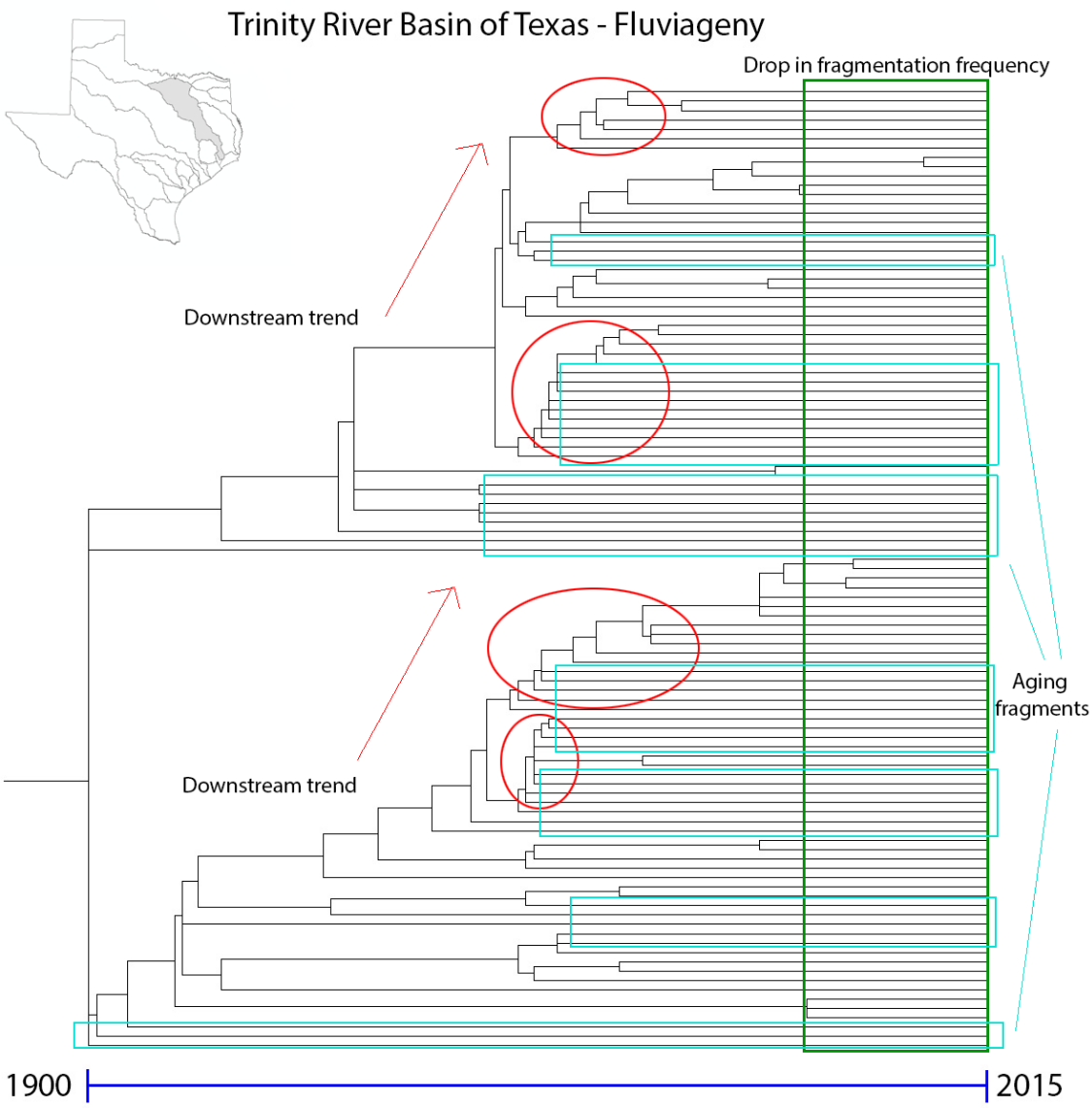

Figure 7. Fluviageny for the Trinity River Basin, using only dams along the uninterrupted primary river network.

We can use these trees to perform quantitative analyses, even without local species data for comparison, to locate patterns in fragmentation. In addition to the degree of fragmentation, N (depth of the node), several other calculations can be computed in order to find comparative fragmentation patterns within and among basins. In order to find two similar patterns of fragmentation, three factors must be considered: N or the depth of the node (the degree of fragmentation, based on the previous damming patterns), the upstream vs downstream relationship of damming (as these may have different effects) and the temporal scale, both where it sits within the chosen temporal range (in the case of the chosen fluviagenies, 1900-2015) and the amount of time that has passed in between fragmentation. For example, within the Trinity River Basin are several visually similar patterns of fragmentation, which could be quantitatively analyzed and compared to detect near-exact matches. The red circles (i.e., staircase patterns) are examples of these similarities as they have similar depths (degrees of fragmentation), exist in the same time-period, and have similar branch-lengths. For intra-basin comparisons, many of the red 'staircase' effects of subsequent downstream damming have similar patterns between fluviagenies. Computationally, these patterns can be detected from browsing the information stored in these trees in tools such as the ETE2 tree browser (Huerta-Cepas, Dopazo, & Gabaldon, 2010), which allows users to view node labels, distances between nodes, and node depth. Let's say we pick two 'staircase' patterns from two different basins. Each staircase contains 7 nodes and exists from 1960-1967. Each node is 1-year apart, and both staircases start at depth 9 and end at depth 16. In this case, these would be considered identical patterns. If we were to map fish phylogenies for a select species or group of fish at these locations to these patterns, assuming ample data is available, we may expect fish to evolve similarly post-fragmentation and for population changes to be similarly affected by the same pattern. We wouldn't expect to see major evolutionary changes in such a small period of time but could predict future speciation.

It's important to consider when using fluviagenies that, while they do point out patterns in hydrological processes, caution must be exerted when using them in consideration of external factors. For example, while two fluviageny subtrees may have near-identical patterns of fragmentation, we may predict that since two guilds of the same fish species reacted similarly to these patterns, they may react the same way to such future patterns. However, there could be other factors influencing speciation or population changes, such as shifts in local climates or non-anthropogenic fragmentation unseen in the fluviageny.

Fish Population Analysis

Figure 8 shows the results of the Colorado River Basin ArcGIS map for large-dam fragmentation from the 10 largest dams in the basin. The fragment of the 1963 dam did not have any cyprinid collecting instances. Several other fragments had limited collecting instances, such as 1 and 2.1.1, which made it difficult to produce reliable analyses on comparisons across the fragmented river. What we lack in this map the ability to distinguish how these fragments existed at certain points in time. Fragment 2.1.2.2.2.2 was once a part of fragments 2, 2.1, 2.1.2, 2.1.2.2, and 2.1.2.2.2. This is easy to understand when looking at the numbers themselves but temporally it is difficult to visualize just by looking at the map. Fluviagenies solve this, as in Figure 9, which shows how easy it is easy to trace this history on a fluviageny tree drawn from the map of the Colorado River Basin.

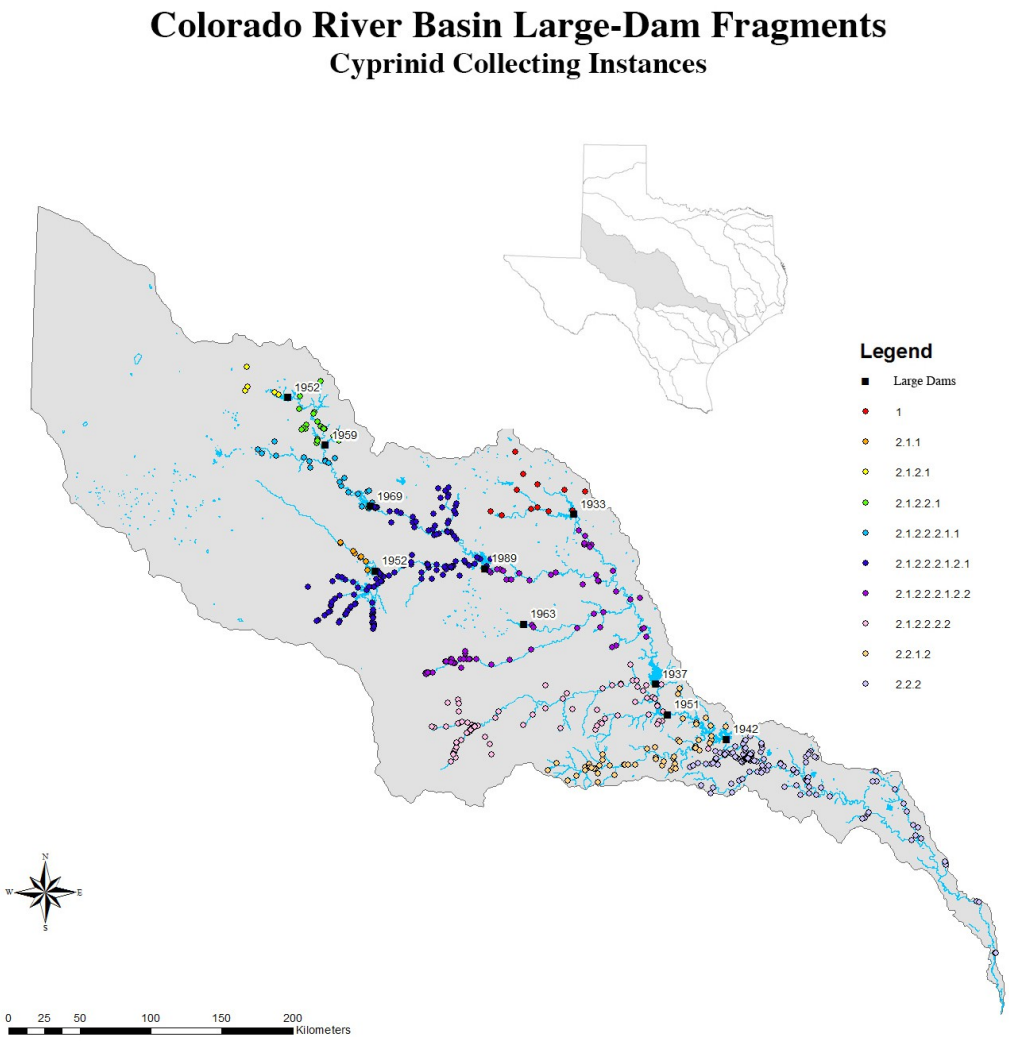

Figure 8. An example map of the Colorado River Basin, split into eleven segments created as a result of the fragmentation caused by the placement of the ten largest dams in the basin. Dots represent Cyprinid fish collecting instances, obtained from Fishes of Texas data (Hendrickson & Cohen, 2010). These are color-coded based on the fragments they are located in. Note that only ten fragments fluviageny codes, with color codes, are shown, since one fragment (upstream of 1963) has no collecting instances.

Figure 9 is a small-scale fluviageny of the fragmentation caused by the major dams. The colored dots were added to represent the shared histories of each fragment and correspond to the colored dots that exist within each fragment in Figure 8. These dots are scaled based on the general size of each fragment with respect to the sizes of the other fragments; the size is based on an estimation of the area of the subbasin created by fragmentation and is not a quantitatively calculated value but a reference for fragment size relationships. The dam construction dates for this tree were resolved so that polytomies did not exist. For example, dams 'e' and 'f' both occurred in 1952 but were approximately 6 months apart, so a gap was created and the polytomy resolved. It would have been more difficult to visualize the temporal split of fragments without the use of this tree, since dams surrounding a fragment may have different construction dates. With the tree, each branch has a spatial location and temporal boundary, determined by the two dams (or nodes) that sit at the ends of the branch. While GIS is used to visualize the spatial aspect, and temporal data can be symbolized to show changes over time, it does not provide the same means of mapping information onto a temporal history, only a spatial plane. In this case, fragment history is written to a tree, which can be used to map fish populations, statistics related to those populations, or biological information, and to perform analyses using the tree from that information. Points in time where major changes in fragmentation or fish populations have occurred can be seen from this tree, as well as unique patterns that otherwise may not have been noticed.

As an example of the potential for fluviagenies to be used for species analysis, species richness numbers were calculated for every fragment at every range of time during fragmentation. The calculations used Menhinick's index for species richness, which is $D = s \div \sqrt{N}$ where s is the number of different species in the sample and n is the total number of organisms in the sample. The resulting richness values were placed onto the tree in Figure 9 in bold, with red values being unreliable due to small sample sizes. The changes in species richness can

be observed by following one of the colored dots, which represents a shared history, along the tree. There were no significant patterns found in the tree as a whole, which may be due to lack of sufficient data across all segments. Generally, most species richness values remained fairly consistent across each fragment, with some dropping steadily. In landscape fragmentation, cross-species generalizations in response to fragmentation may not accurately represent general changes and may require individual species analysis comparisons (Betts, et al., 2014), which might be the same in riverine fragmentation. Cross-species comparison studies could be facilitated by mapping species onto fragment trees and comparing them to one another. Such studies emphasize the usefulness of large-scale fluviageny trees, such as those in figures 5, 6, and 7, which, when combined with sufficient data for individual species of fish, could draw out patterns previously undetected or help predict future fragmentation effects.

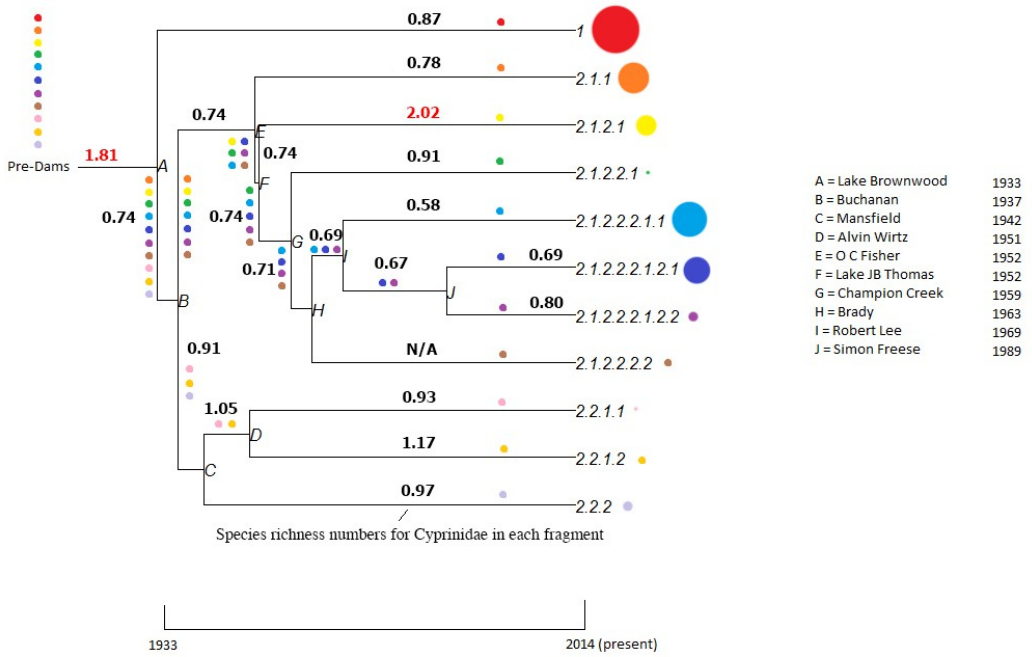

Figure 9. A fluviageny tree corresponding to Figure 8. This tree shows the usefulness in the visualization of temporal data even for small datasets such as the ten large dams. Histories of fish collecting instance, represented as colored dots corresponding to the fragments in Figure 8, can be traced back through the tree to see what fish were part of which fragments at what points in time. Dots are scaled based on the sizes of the fragments relative to one another. Numbers drawn on branches represent Menhinick's index for species richness. Red values represent species richness indices that did not have ample data for calculation and may be unreliable.

DISCUSSION

Benefits of Fluviagenies

There are many benefits to using fluviagenies for biodiversity research. The most immediately apparent benefit is that fluviagenies provide visualization on a temporal scale that is not possible in geographic information systems. When viewing a fragmented river system in a geographic map, it is easy to understand the visual representation of fragmentation of the river system at one point in time, such as the present day. All points existing on the map, representing dams, show us where barriers may create blocks to river flow. In between these barriers are sections of the river that have ecosystems that may be affected by the fragmentation and thus may evolve independently of one another. However, it's important to note when these barriers were created, for how long these fragments have existed, and how they are related to other fragments. It's impossible to visualize this on a 2-dimensional map and would be extremely difficult to understand in three-dimensional visualization. The fluviageny transfers the history of fragmentation into a readable visual that shows how each fragment is related to one another. The history of one fragment can be traced back through time by starting at the leaf node (tip) representing a fragment and moving back to each ancestral parent fragment. Tracing through fluviagenies allows us to see the relationships between fragments at any point in time and not just a snapshot at a single point in time as shown on a map. Subtrees can also be extracted from full fluviageny trees to analyze and compare select timeframes or histories of individual fragments.

Fluviagenies are easy to build with knowledge of geographic information systems and programming. Once fluviagenies are created, they can be exported into Newick format and provide means of analyzing fragmentation data across time that were never available before, including running this data through phylogeny tools. Due to the massive number of phylogenetic toolsets that exist in various interfaces and programming languages, once a fluviageny is constructed,

there are limitless ways to manipulate and analyze the tree, including comparing fluviagenies from different areas to one another, tracing species changes over time onto fluviagenies, or even mapping species phylogenetic trees onto fluviagenies. Any number of phylogenetic metrics can be applied to fluviagenies, such as Pagel's lambda and distinctiveness, which were used for analyzing terragenetic trees (Ewers, et al., 2013). This opens up many possible future uses for fluviagenies.

Future Uses for Fluviagenies

The next step for fluviagenies is to put them to use. There are several possible scenarios where fluviagenies could be applied to phylogenetic and biodiversity research. They can be compared against null models (i.e., a continuous and uninterrupted river system) to determine the impact of fragmentation on species changes and to predict future changes in biodiversity patterns. Fragment length is important, as certain fish populations, such as pelagic spawners, require a certain length of uninterrupted river reaches in order for their eggs or larvae to drift long enough to allow them to hatch. While these lengths could be studied easily outside of fluviagenies, it's with relation to time, however, that is important to us and is aided by fluviageny trees, as different periods of isolation time (i.e., the age of a fragment) may have similar effects in comparable ecosystems. Fluviagenies with similar patterns of fragmentation could also be compared against one another. If similar patterns caused comparable effects on the same species, it could be predicted that such patterns would lead to similar effects in other areas. Different guilds of fishes could be hypothesized to respond differently to certain patterns of fragmentation. Knowledge of historical habitat fragmentation could help us predict how anthropogenic fragmentation might affect future populations via phylogeographic analysis. Such information is vital to understanding future damming practices and even the environmental effects of removing current

dams, which is becoming a prominent issue in recent times (Babbitt, 2002).

Similarly to Ewers et al.'s research using terragenies, variance can be used to analyze habitat loss and predict possible rates of extinction, especially as a result of fragmentation. Communities between different fragments can be compared to predict shared species between communities. Ewers et al.'s found that community similarity declined with terragenetic distance, meaning the more fragmented the landscape became, the less similar more distant habitats became. The authors also used terragenies to predict expected proportions of endemic species, which is vital to conservation research (Ewers, et al., 2013). Fluviagenies could potentially be used to perform all the types of analyses conducted using terragenies. Terragenetic models have been validated to work effectively for biodiversity research and since fluviagenies are strongly based on terragenies there is already existing evidence that such tools are useful.

It's important to note that while fluviagenies work well for analyzing fragmentation patterns, there are a number of extraneous variables that could interfere with conclusions drawn from fluviageny-based research. There are considerations to make for the level of fragmentation caused by damming, such as dam size and type or features of dams or basins that may allow fish to pass over, through, or around the dams. Not all of this data is recorded, even in the most comprehensive available dam datasets. There are also other causes of fragmentation such as natural barriers or drought. Any predictions made about fish population changes may also be affected by additional variables such as disease or widespread ecosystem changes outside of barrier-based changes. However, these factors would exist whether the research was conducted using fluviagenies or not, so this tool does not create additional problems that would not have otherwise existed, it simply provides researchers with additional ways to manipulate and visualize fragmentation data.

Applying Fluviageny Concepts to Other Disciplines

Fluviagenies, as a concept, can help us think about using phylogenetic methods for other, even more obscure applications. The concept of the fluviageny is based on the phylogenetic ideas of splits over time, i.e., evolution of species and ancestral relationships between those species. Any form of data that takes on similar attributes of the data used in fluviagenies could benefit from phylogenetic toolsets. As long as the data contains a temporal aspect and historic splits with relationships between previous ancestors, it could be generated as a phylogenetic tree. While non-biological fields would not need to be processed through real phylogenetic analysis, the tools that allow manipulation of trees in Newick format, for example, could be used for analysis in other areas of interest. One possible use would be tracking the history of a large company over time. For example, a retail chain may start out as a small store but expand into other geographic areas by opening other stores. The corporate chain my split over time as well. However, this may also include merges and not just splits, but the use of trees to demonstrate temporal history would still be useful as a visualization tool. Another similar scenario is cadastral mapping, which displays changes in property boundaries over time in fixed geographic areas and includes splits and merges in property. Phylogenetic toolsets could be used to analyze the relationships between lineages of the properties. These tools have many potential applications outside of their original intended use and in the future will only continue to grow in power.

APPENDIX

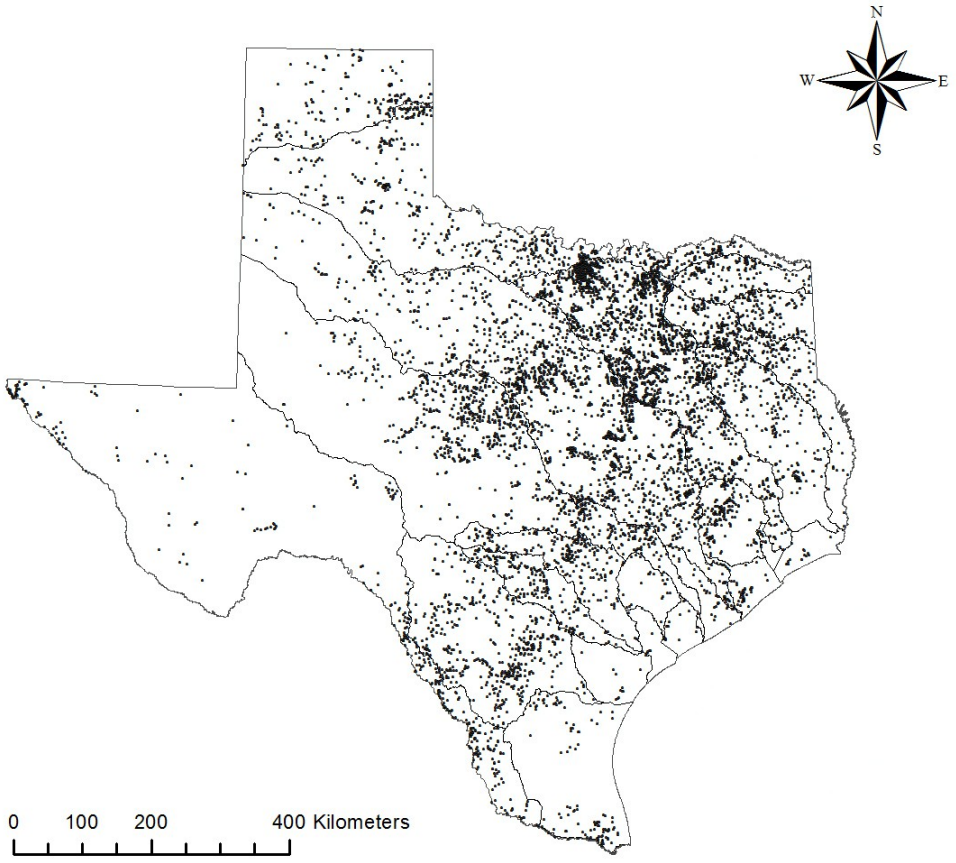

Figure A1. Dams in the State of Texas (n=3,175), using Texas Commission on Environment Quality (TCEQ, 2013) data, shown with all primary drainage basins in Texas.

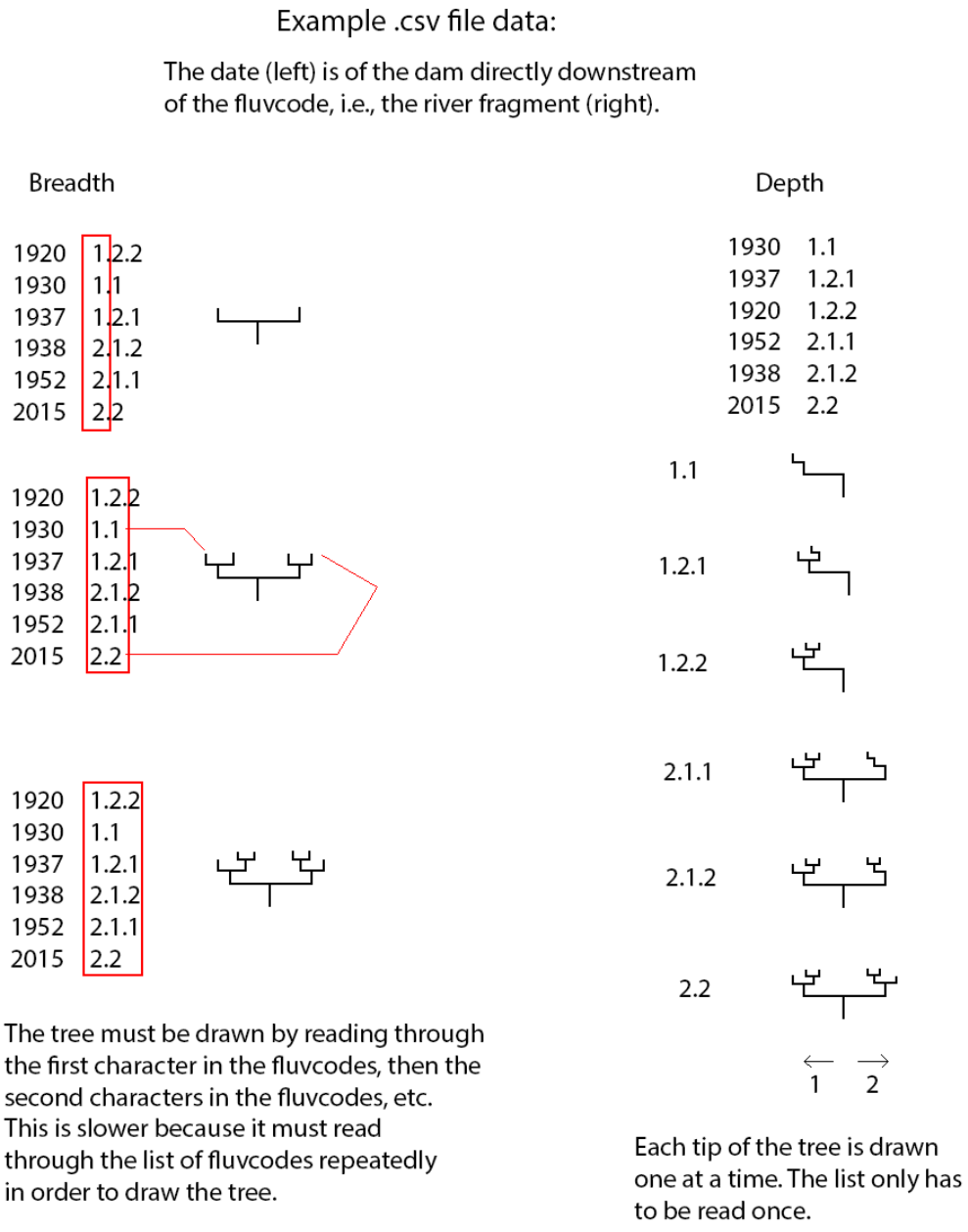

Figure A2. Two methods of drawing fluviageny trees from .csv files: breadth-first and depth-first. The depth-first method is the preferred method and the one that was coded for use in this study.

```
##imports csv with 2 columns based on user-specified path: 1-dates; 2-fluvcodes; creates list from csv
import sys
import os
import csv
user_input = raw_input("Please enter the full filepath to the .csv with no quotations: ")
assert os.path.exists(user_input), "File not found at the specified path: , "+str(user_input)
with open(user_input) as infile:
     reader = csv.reader(infile)
     list1=list()
     for row in reader:
          row="".join(row)
          list1.append(row)

##create blank tree; sort the list by dates first, then read first date and assign it to root of tree
##add two splits for first tree - "1" and "2"; then sort list by fluvcode
from ete2 import Tree
list1.sort()
first = list1[0]
t=Tree(name=first[0:4])
t.add_child(name="1")
t.add_child(name="2")
list1=sorted(list1, key=lambda list1: (list1[4:]))

##go through list and for each fluvcode, place it appropriately into the tree (depth-first)
##n is for index of characters in fluv code, to read through the series of 1's and 2's and make children for each
for fluv in list1:
     n=7
     while n <= len(fluv):
          if not t.search_nodes(name=fluv[4:n]):
               A = t.search_nodes(name=fluv[4:n-2])[0]
               A.add_child(name=fluv[4:n])
          n=n+2

##go back through the completed tree, replacing internal nodes (non-leaf fluvcodes) with dates
##z and trunc read through each fluvcode ignoring the first four characters (dates)
##this algorithm relies on the fact that each date represents the next downstream dam,
##so the next time a '1' (1=upstream) appears in an internal fluv, that fluv segment is upstream of the next dam date
##when reading from end of fluvcode to beginning, the first '1' encountered serves as a signal
##the '1' shows that the remaining fluvcode (before the 1) represents the upstream segment and is replaced with dam name
for fluv in list1:
     z=len(fluv)-1
     while z >=4:
          if fluv[z] == '1':
               trunc = z - 1
               z=0
          else:
               z = z-1
     dam = fluv[4:trunc]
     if t.search_nodes(name=dam):
          A = t.search_nodes(name=dam)[0]
          A.name = fluv[0:4]

##scale all nodes in trees based on difference in dam dates; leaf nodes are assigned value of user specified year
user_year = raw_input("Enter desired end year for tree, e.g., current year: ")
m=t.traverse()
for node in m:
     if node.is_leaf():
          x=int(user_year)
          c=node.up
          y=int(c.name)
          z=x-y
          node.dist=str(z)
     elif node.is_root():
          node.dist='10'
     else:
          x=int(node.name)
          c=node.up
          y=int(c.name)
          z=x-y
          node.dist=str(z)

##view tree in ETE tree browser, which gives the option of browsing and exporting into Newick for other applications
t.show()
```

Figure A3. Python code that draws fluviagenies using data exported from ArcGIS.

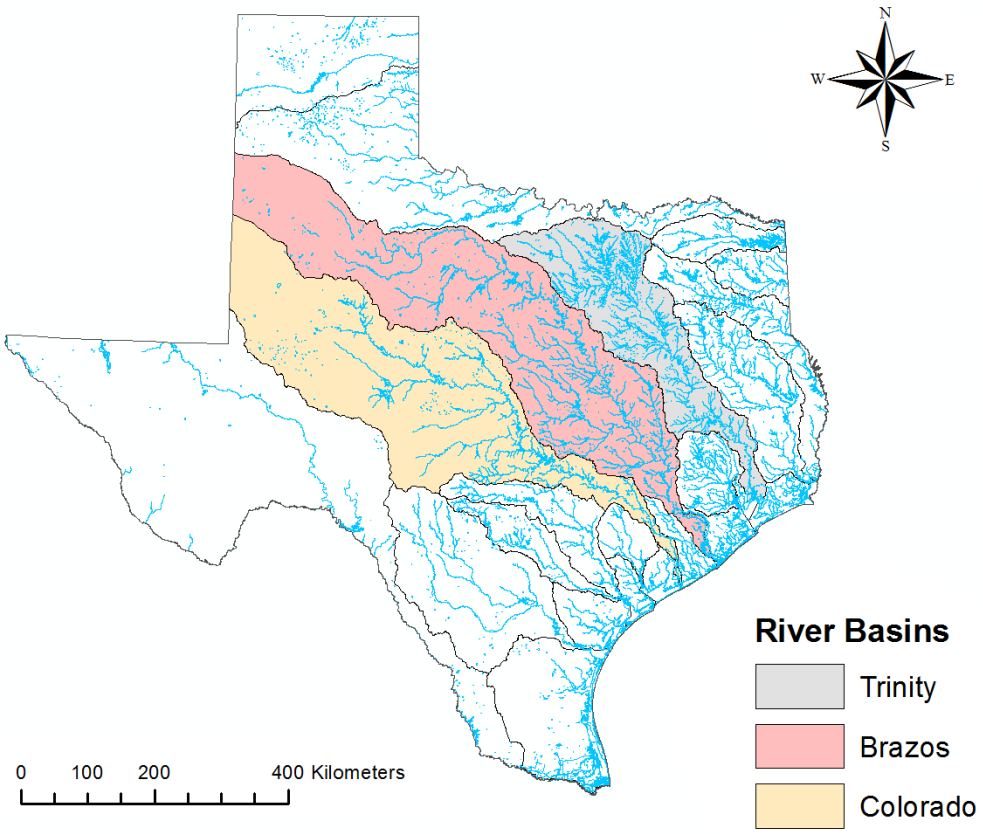

Figure A4. A map of the Texas river network with basins outlined throughout the state. The Trinity, Brazos, and Colorado River basins were selected as test basins for fluviageny tree generation and analysis.

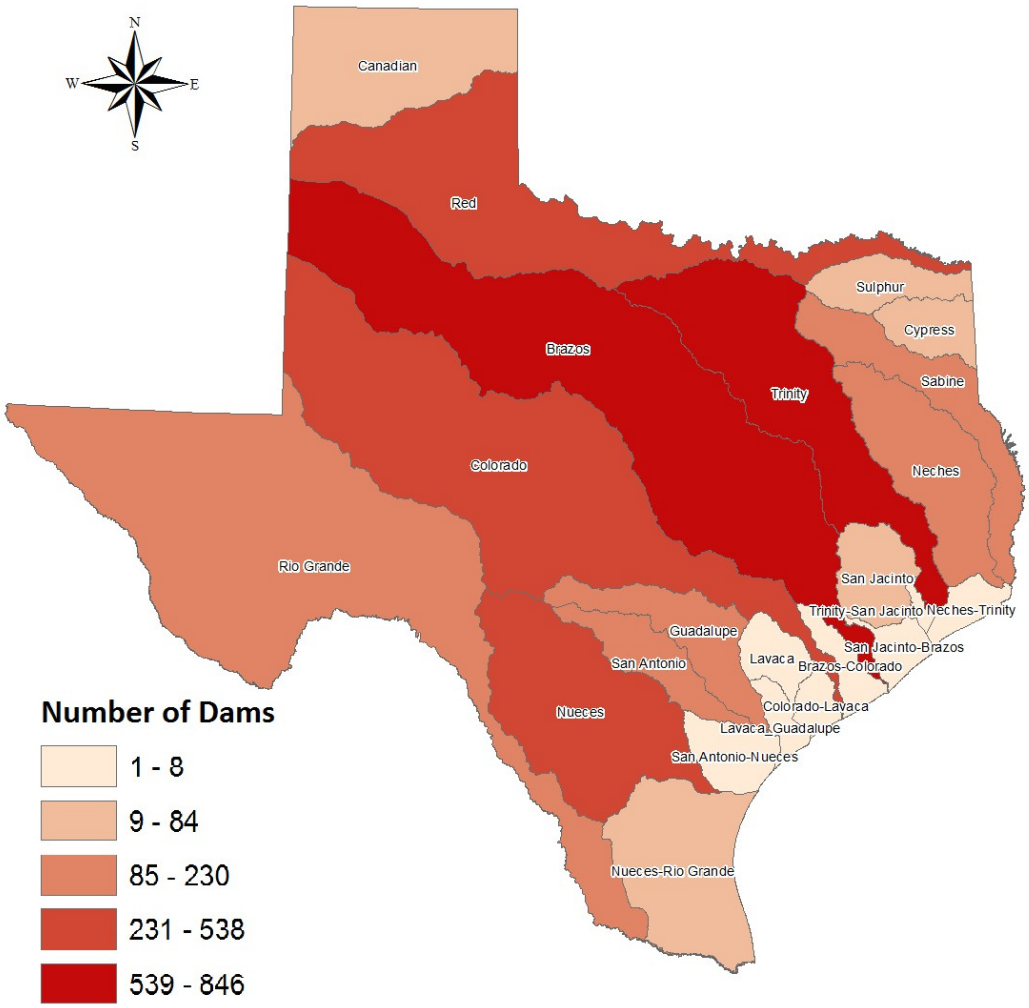

Figure A5. Dam counts for river basins in the state of Texas using data from the Texas Commission on Environmental Quality (TCEQ, 2009).

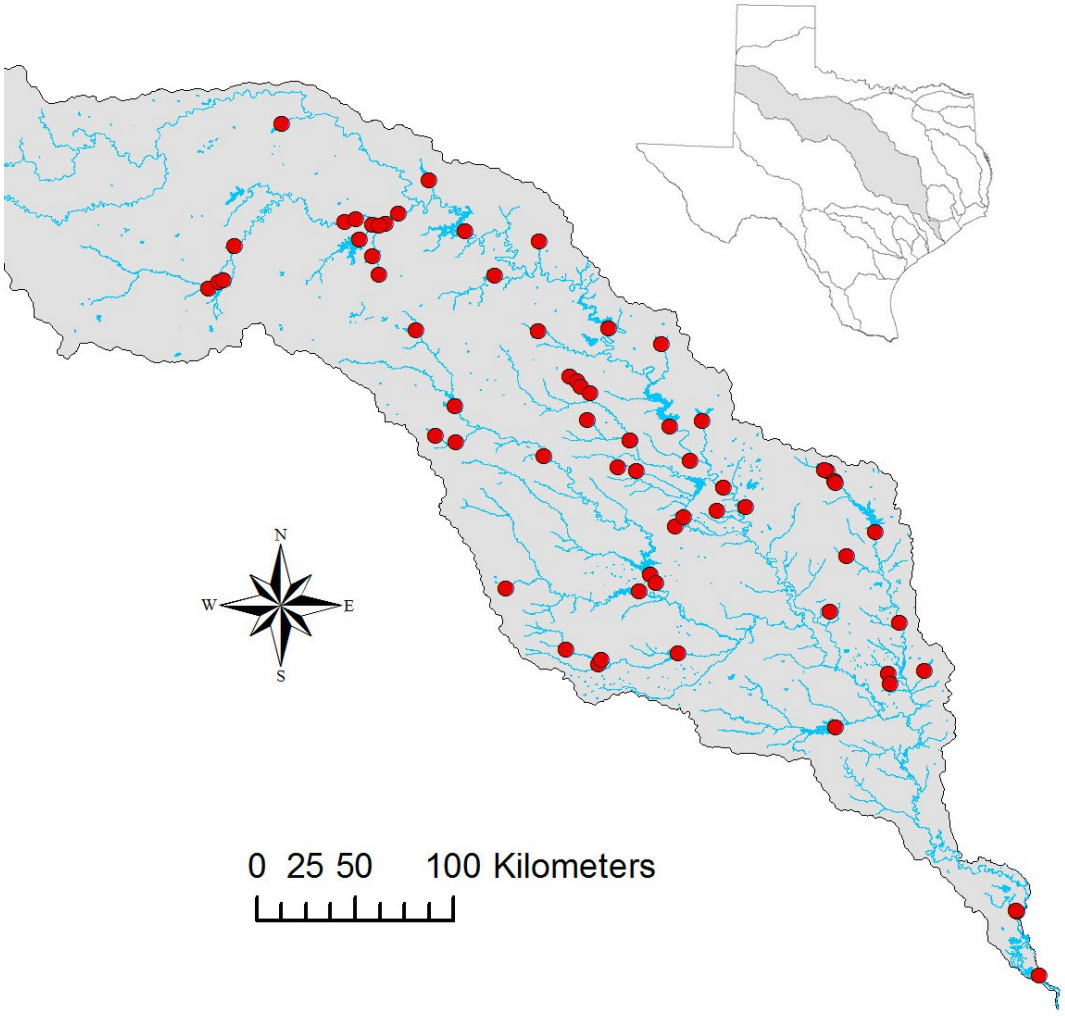

Figure A6. Map of the dams selected for the Brazos River Basin fluviageny (n=67). Dams are from the National Anthropogenic Barriers Dataset (Ostroff, Wieferich, Cooper, & Infante, 2012) and were selected only if they were connected to the main river network.

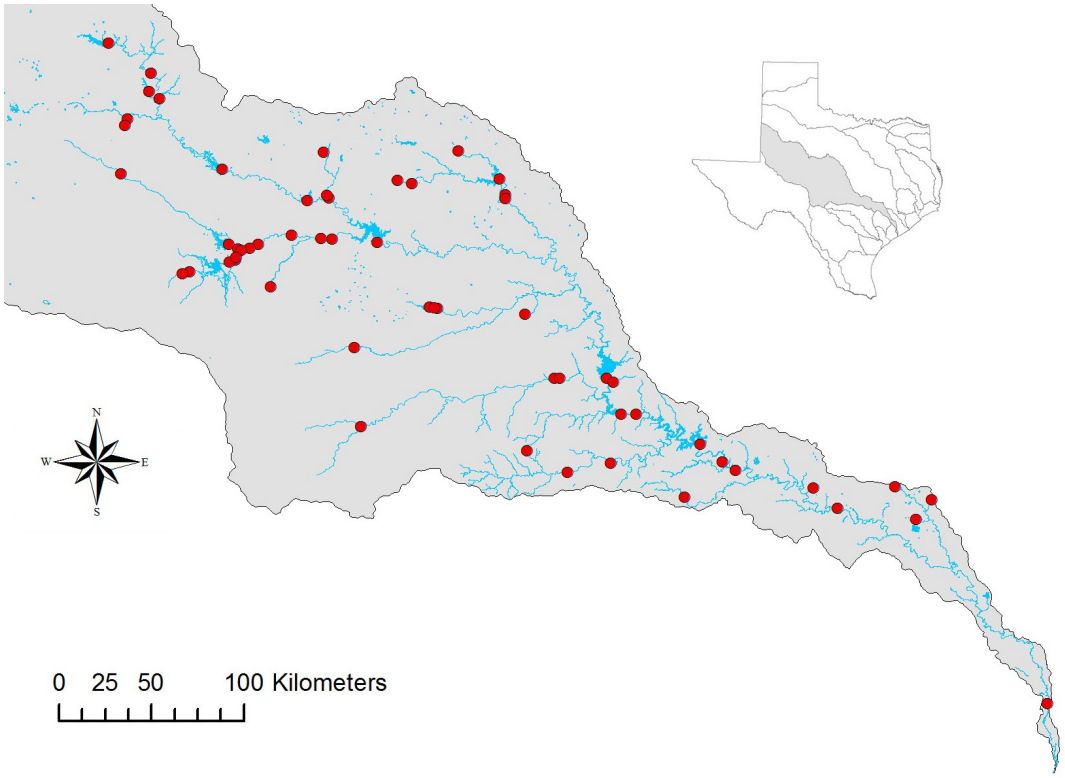

Figure A7. Map of the dams selected for the Colorado River Basin fluviageny (n=60). Dams are from the National Anthropogenic Barriers Dataset (Ostroff, Wieferich, Cooper, & Infante, 2012) and were selected only if they were connected to the main river network.

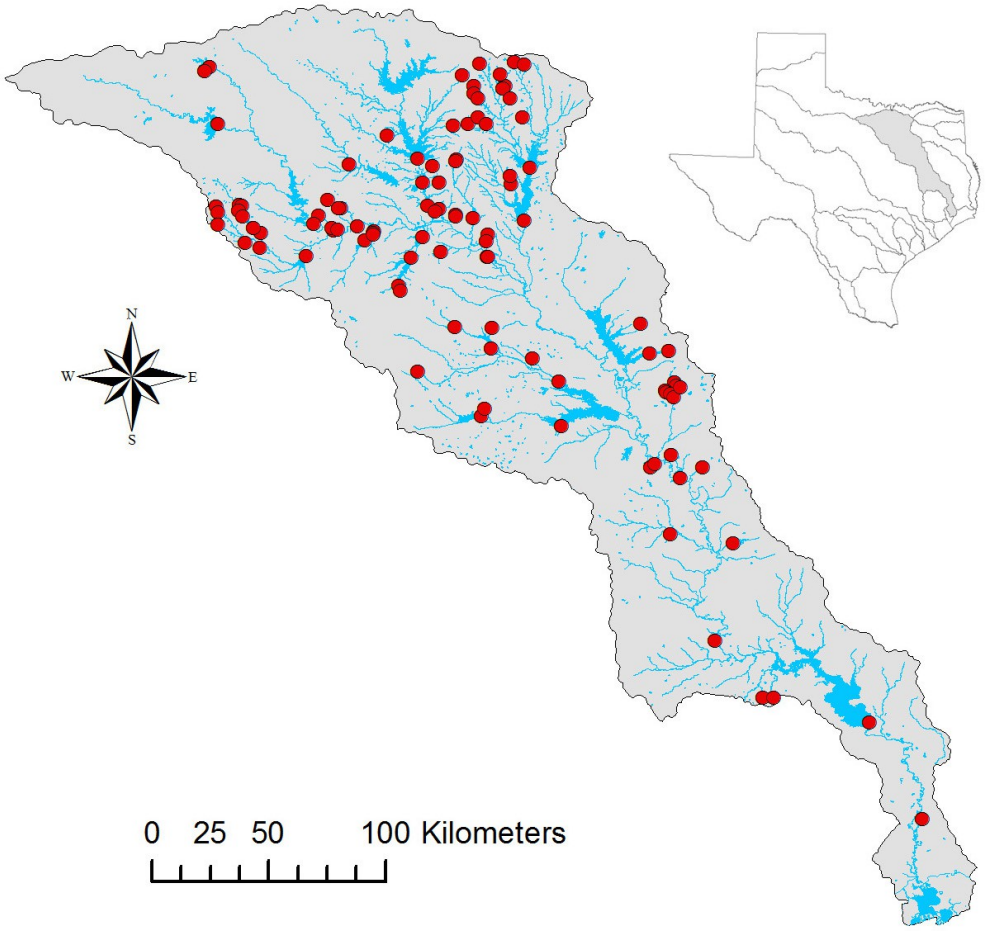

Figure A8. Map of the dams selected for the Trinity River Basin fluviageny (n=102). Dams are from the National Anthropogenic Barriers Dataset (Ostroff, Wieferich, Cooper, & Infante, 2012) and were selected only if they were connected to the main river network.

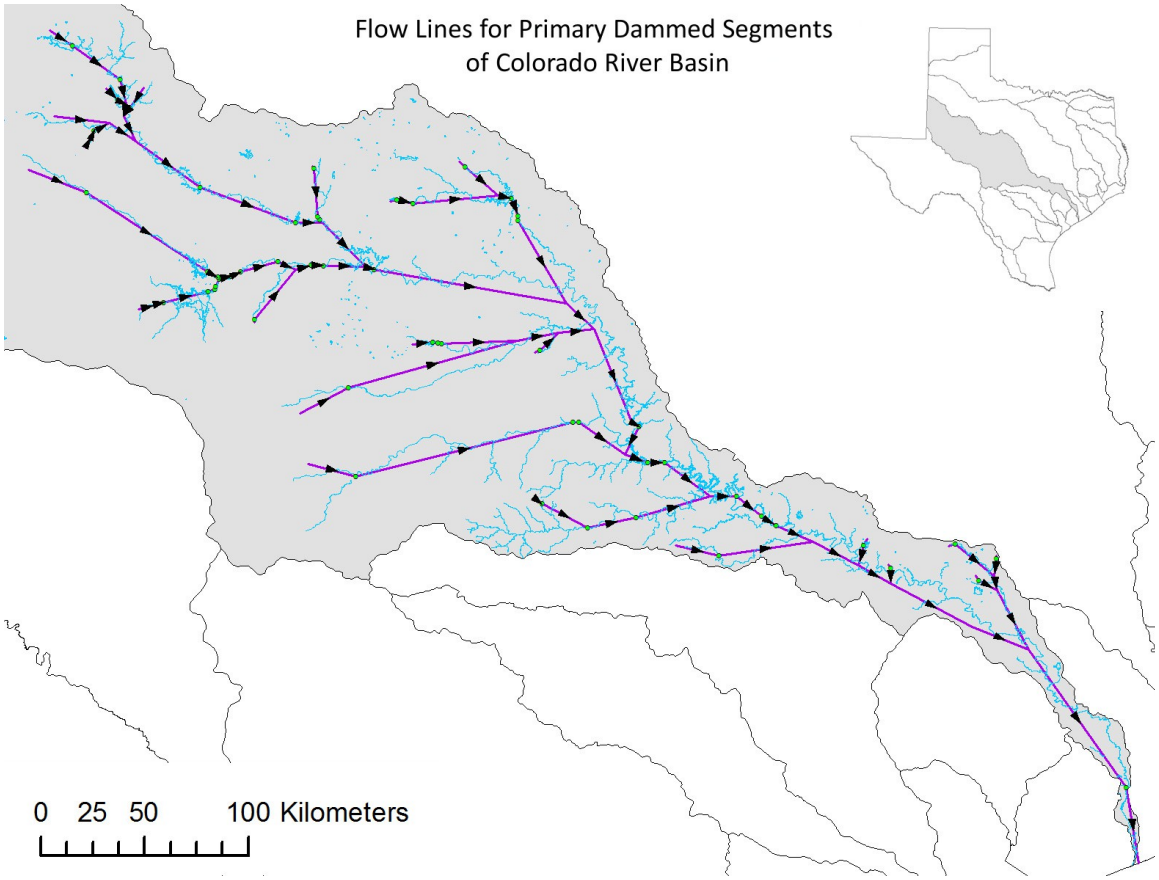

Figure A9. Simplified flow direction of selected dam-bound fragments in the Colorado River Basin of Texas.

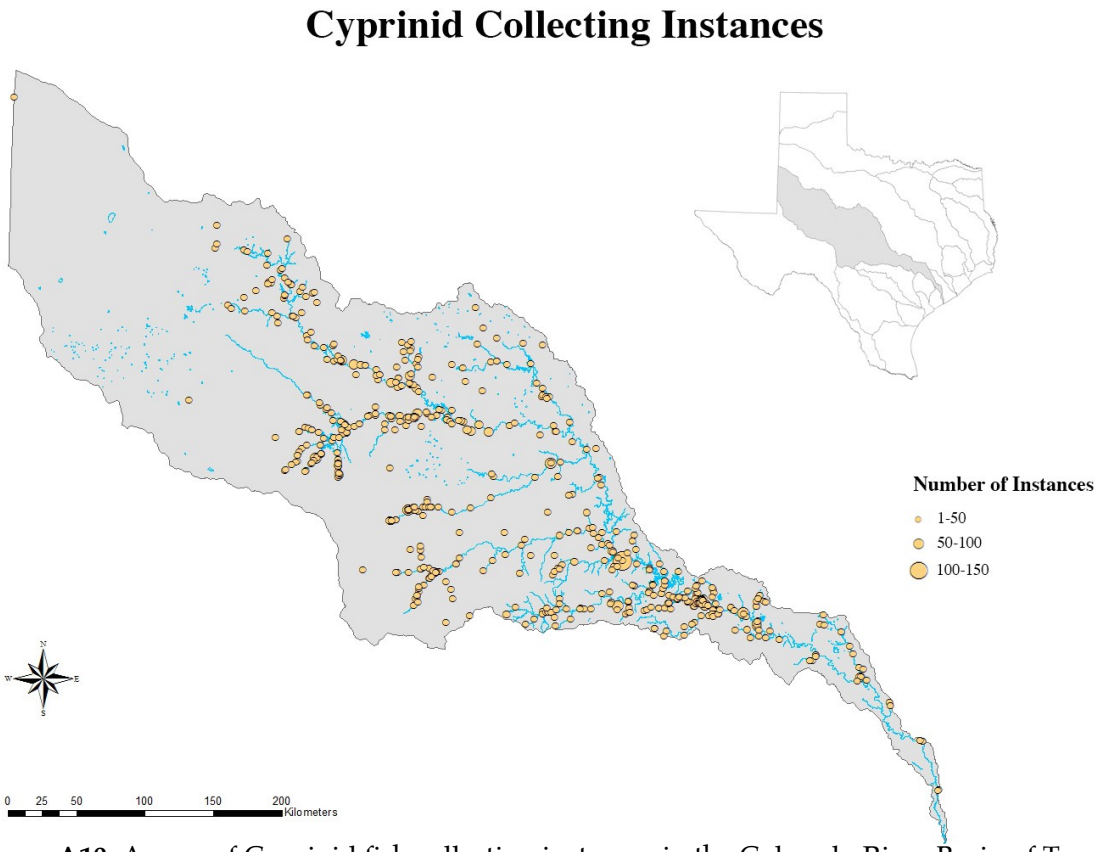

Figure A10. A map of Cyprinid fish collecting instances in the Colorado River Basin of Texas (n=3,191). Stacked dots, i.e., fish collected at the same location, are aggregated into larger, scaled dots. Data obtained from the Fishes of Texas database (Hendrickson & Cohen, 2010).

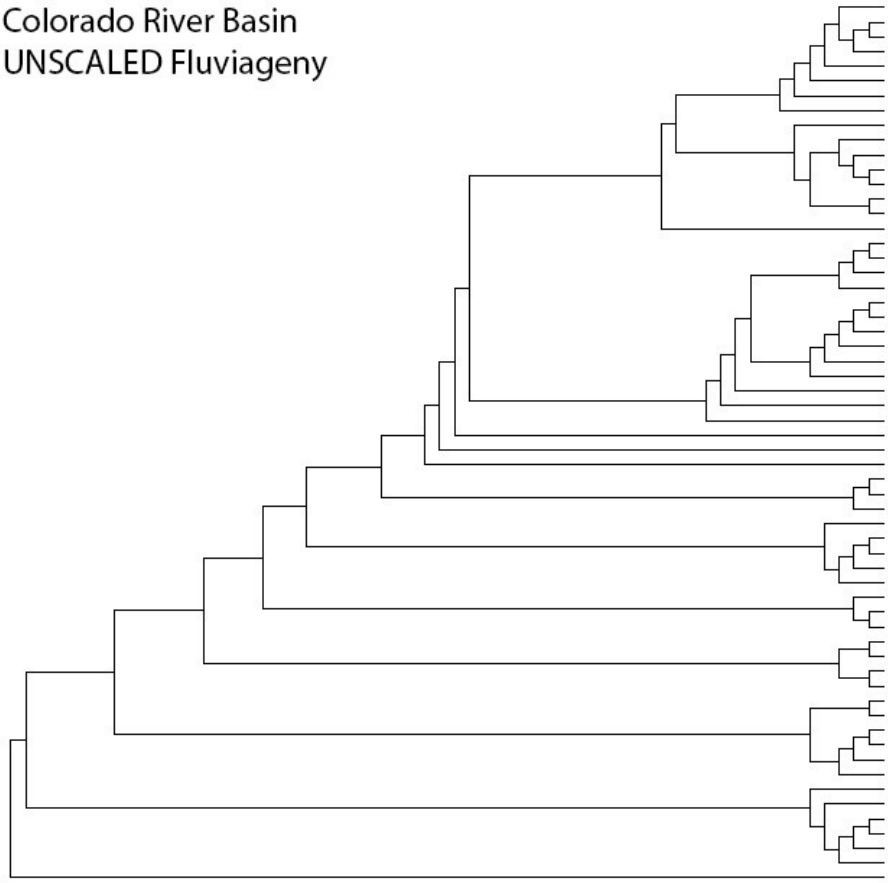

Figure A11. An example of an unscaled fluviageny tree for the Colorado River Basin. Each individual occurrence of fragmentation, per dam (node), can be seen in the unscaled version. The temporally scaled fluviagenies compress nodes that were created at the same time (according to the data, i.e., same year) into unresolved polytomies.

REFERENCES

Association of State Dam Safety Officials. (2010). *State Dam Safety Classification Schemes*. Retrieved from http://www.damsafety.org

Babbitt, B. (2002). What goes up, may come down. *BioScience*, *52*(8), 656–658. http://doi.org/10.1641/0006-3568(2002)052[0656:WGUMCD]2.0.CO;2

Betts, M. G., Fahrig, L., Hadley, A. S., Halstead, K. E., Bowman, J., Robinson, W. D., … Lindenmayer, D. B. (2014). A species-centered approach for uncovering generalities in organism responses to habitat loss and fragmentation. *Ecography*, *37*, 517–527. http://doi.org/10.1111/ecog.00740

Bowman, M. B. (2002). Legal perspectives on dam removal. *BioScience*, *52*(8), 739–747. http://doi.org/10.1641/0006-3568(2002)052[0739:LPODR]2.0

Cavender-Bares, J., Ackerly, D. D., & Kozak, K. H. (2012). Integrating ecology and phylogenetics: the footprint of history in modern-day communities. *Ecology*, *93*, S1–S3. http://doi.org/10.1890/12-0092.1

Chin, A., Laurencio, L. R., & Martinez, A. E. (2008). The Hydrologic importance of small- and medium-sized dams: examples from Texas. *The Professional Geographer*, 238–251. http://doi.org/10.1080/00330120701836261

Chisham, B., Wright, B., Le, T., Tran, C. S., & Pontelli, E. (2011). CDAO-Store: ontology-driven data integration for phylogenetic analysis. *BMC Bioinformatics*, *12*(98). http://doi.org/10.1186/1471-2105-12-98

Collier, M., Webb, R. H., & Schmidt, J. C. (1996). *Dams and Rivers: A Primer on the Downstream Effects of Dams*. Arizona: U.S. Geological Survey Circular 1126.

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., & Peterson, A. T. (2010). VertNet: a new model for biodiversity data sharing. *PLoS Biology*, *8*(2), e1000309. http://doi.org/10.1371/journal.pbio.1000309

Cooper, A. R. (2013). *Effects of dams on streams of the conterminous United States: characterizing patterns in habitat fragmentation nationally and fluvial fish response in the Midwest* (Master's Thesis). Michigan State University. Retrieved

from http://etd.lib.msu.edu/islandora/object/etd%3A2191

Dornelas, M., Magurran, A. E., Buckland, S. T., Chao, A., Chazdon, R. L., Colwell, R. K., … Vellend, M. (2013). Quantifying temporal change in biodiversity: challenges and opportunities. *Proceedings of the Royal Society of B: Biological Sciences*, *280*(1750). http://doi.org/10.1098/rspb.2012.1931

Dynesius, M., & Nilsso, C. (1994). Fragmentation and flow regulation of river systems in the northern third of the world. *Science*, *266*(5186), 753–761.

Ehrenberg, J. E., & Steig, T. W. (2003). Imroved techniques for studying the temporal and spatial behavior of fish in a fixed location. *Journal of Marine Science*, *60*, 700–706. http://doi.org/10.1016/S1054–3139(03)00087-0

ESRI (Environmental Systems Resource Institute). (2013). *ArcMap 10.1*. Redlands, California: ESRI.

Ewers, R. M., Didham, R. K., Pearse, W. D., Lefebvre, V., Rosa, I. M. D., Carreiras, J. M. B., … Reuman, D. C. (2013). Using landscape history to predict biodiversity patterns in fragmented landscapes. *Ecology Letters*, *16*, 1221–1233. http://doi.org/10.1111/ele.12160

Fausch, K. D., Torgersen, C. E., Baxter, C. V., & Li, H. W. (2002). Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *BioScience*, *52*(6), 483–498. http://doi.org/10.1641/0006-3568(2002)052[0483:LTRBTG]2.0.CO;2

Fukushima, M., Kameyama, S., Kaneko, M., Nakao, K., & Steel, A. E. (2007). Modelling the effects of dams on freshwater fish distributions in Hokkaido, Japan. *Freshwater Biology*, *52*, 1511–1524. http://doi.org/10.1111/j.1365-2427.2007.01783.x

Goloboff, P. A., Farris, J. S., & Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics*, *24*, 774–786. http://doi.org/10.1111/j.1096-0031.2008.00217.x

Goodall, J., Maidment, D., & Sorenson, J. (2004). Representation of spatial and temporal

data in ArcGIS. Presented at the AWRA GIS and Water Resources Conference III, Nashville, TN, USA.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, *19*(9), 497–503. http://doi.org/10.1016/j.tree.2004.07.006

Graph, W. L. (2005). Geomorphology and American dams: the scientific, social and economic context. *Geomorphology*, *71*, 3–26. http://doi.org/10.1016/j.geomorph.2004.05.005

Greathouse, E. A., Pringle, C. M., McDowell, W. H., & Holmquist, J. G. (2006). Indirect upstream effects of dams: consequences of migratory consumer extirpation in Puerto Rico. *Ecological Applications*, *16*(1), 339–352. http://doi.org/10.1890/05-0243

Gregory, S., Li, H., & Li, J. (2002). The conceptual basis for ecological responses to dam removal. *BioScience*, *52*(8), 713–723. http://doi.org/10.1641/0006-3568(2002)052[0713:TCBFER]2.0.CO;2

Grytness, J.-A., & Romdal, T. S. (2008). Using museum collections to estimate diversity patterns along geographical gradients. *Folia Geobot*, *43*, 357–359. http://doi.org/10.1007/s12224-008-9017-6

Guralnick, R., & Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global diversity patterns and processes. *Bioinformatics*, *25*, 421–428. http://doi.org/10.1093/bioinformatics/btn659

Hall, C. J., Jordaan, A., & Frisk, M. G. (2011). The historic influence of dams on diadromous fish habitat with a focus on river herring and hydrologic longitudinal connectivity. *Landscape Ecology*, *26*, 95–107. http://doi.org/10.1007/s10980-010-9539-1

Hart, D. D., Johnson, T. E., Bushaw-Newton, K. L., Horwitz, R. J., Bednarek, A. T., Charles, D. F., … Velinsky, D. J. (2002). Dam removal: challenges and

opportunities for ecological research and river restoration. *BioScience*, *52*(8), 667–681. http://doi.org/10.1641/0006-3568(2002)052[0669:DRCAOF]2.0.CO;2

Hart, D. D., & Poff, N. L. (2002). A special section on dam removal and river restoration. *BioScience*, *52*(8), 653–655. http://doi.org/10.1641/0006-3568(2002)052[0653:ASSODR]2.0.CO;2

Hendrickson, D. A., & Cohen, A. E. (2010). *Fishes of Texas Project and Online Database*. Austin, Texas: Texas Natural History Collection, a division of Texas Natural Science Center, University of Texas. Retrieved from www.fishesoftexas.org

Hill, A., & Guralnick, R. (2008). A case study for distributed systems and automated biodiversity informatics: genomic analysis and geographic visualization of disease evolution (pp. 270–279). Presented at the British National Conference on Databases 5071, Birmingham, UK.

Hoyle-Dodson, G. (2009, May). *Integration of ArcGIS hydrology modeling and geoprocessing tools with dam breach flood analysis*. Presented at the 2009 Washington GIS Conference, Bellevue, Washington.

Huerta-Cepas, J., Dopazo, J., & Gabaldon, T. (2010). ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, *11*(24).

Jager, H. I., Chandler, J. A., Lepla, K. B., & Winkle, W. V. (2001). A theoretical study of river fragmentation by dams and its effects on white sturgeon populations. *Environmental Biology of Fishes*, *60*, 347–361. http://doi.org/10.1023/A:1011036127663

Jelks, Howard L., Walsh, S. J., Burkhead, N. M., Contreras-Balderas, S., Diaz-Pardo, E., Dean A. Hendrickson, … Warren. (2008). Conservation status of imperiled North American freshwater and diadromous fishes. *Fisheries*, *33*(8), 372–407. http://doi.org/10.1577/1548-8446-33.8.372

Johnson, S. E., & Graber, B. E. (n.d.). Enlisting the social sciences in decisions about dam removal. *BioScience*, *52*(8), 731–738. http://doi.org/10.1641/0006-

3568(2002)052[0731:ETSSID]2.0.CO;2

Lapp, H., Morris, R. A., Catapano, T., Hobern, D., & Morrison, N. (2011). Organizing our knowledge of biodiversity. *Bulletin of the American Society for Information Science and Technology*, *37*, 38–42.

Ligon, F. K., Dietrich, W. E., & Trush, W. J. (1995). Downstream ecological effects of dams. *BioScience*, *45*(3), 183–192.

Lindenmayer, D. B., Welsh, A., Blanchard, W., Tennant, P., & Donnelly, C. (2014). Exploring co-occurrence of closely-related guild members in a fragmented landscape subject to rapid transformation. *Ecography*, *37*, 1–10. http://doi.org/10.1111/ecog.00939

Maidment, D. R. (2003). *Arc Hydro: GIS for Water Resources*. Redlands, California: ESRI Press.

Mamoulis, N., Cao, H., Kolliol, G., Hadjieleftheriou, M., Tao, Y., & Cheung, D. W. (2004). Mining, indexing, and querying historical spatiotemporal data. Presented at the KDD'04, Seattle, WA, USA: ACM.

Matthews, W. J., & Marsh-Matthews, E. (2015). Comparison of historical and recent fish distribution patterns in Oklahoma and Western Arkansas. *Copeia*, *103*(1), 170–180. http://doi.org/10.1643/CE-14-005

McGovern, A., Hiers, N. C., Collier, M., Gagne II, D. J., & Brown, R. A. (2008). Spatiotemporal relational probability trees: an introduction. Presented at the Eighth IEEE International Conference on Data MIning, Pisa, Italy. http://doi.org/10.1109/ICDM.2008.134

Miller, J. T., & Jolley-Rogers, G. (2014). Correcting the disconnect between phylogenetics and biodiversity informatics. *Zootaxa*, *3754*(2). http://doi.org/http://dx.doi.org/10.11646/zootaxa.3754.2.8

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., & O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature*, *409*, 614–618.

Nascimento, M., & Silva, J. (1998). Towards historical R-trees. In *Proceedings of the 1998 ACM Symposium on Applied Computing* (pp. 235–240). Atlanta, GA, USA: ACM. http://doi.org/10.1145/330560.330692

Noel, G., Servigne, S., & Laurini, R. (2005). The po-tree: a real-time spatiotemporal data indexing structure. *Developments in Spatial Data Handling*, 259–270. http://doi.org/10.1007/3-540-26772-7_20

Ostroff, A., Wieferich, D., Cooper, A., & Infante, D. (2013). *2012 National Anthropogenic Barrier Dataset (NABD)*. Retrieved from https://www.sciencebase.gov/catalog/item/512cf142e4b0855fde669828

Page, R. D. M. (2012). Space, time, form: viewing the Tree of Life. *Trends in Ecology and Evolution*, *27*, 113–120. http://doi.org/10.1016/j.tree.2011.12.002

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290.

Parr, C. S., Guralnick, R., Cellinese, N., & Page, R. D. M. (2012). Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology and Evolution*, *27*(2), 94–103. http://doi.org/10.1016/j.tree.2011.11.001

Penev, L., Roberts, D., Smith, V., Agosti, D., & Erwin, T. (2010). Taxonomy shifts up a gear: new publishing tools to accelerate biodiversity research. *ZooKeys*, *50*, i–iv. http://doi.org/10.3897/zookeys.50.543

Perkin, J. S., & Gido, K. B. (2011). Stream fragmentation thresholds for a reproductive guild of Great Plains fishes. *Fisheries*, *36*(8), 371–383. http://doi.org/10.1080/03632415.2011.597666

Perkin, J. S., Gido, K. B., Cooper, A. R., Turner, T. F., Osborne, M. J., Johnson, E. R., & Mayes, K. B. (2014). *The last of the large fragments: how dams and dewatering have transformed Great Plains stream fish communities*. Unpublished paper.

Peuquet, D. J. (2001). Making Space for Time: Issues in Space-Time Data Representation. *Geoinformatica*, *5*(1), 11–32.

Peuquet, D. J., & Duan, N. (1995). An event-based spatiotemporal data model (ESTDM)

for temporal analysis of geographical data. *International Journal of Geographical Information Systems*, *9*(1), 7–24. http://doi.org/10.1080/02693799508902022

Pigot, A. L., & Etienne, R. S. (2015). A new dynamic null model for phylogenetic community structure. *Ecology Letters*, *18*, 153–163. http://doi.org/10.1111/ele.12395

Piwowar, H. A., Becich, M. J., Bilofsky, H., & Crowley, R. S. (2008). Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Medicine*, *5*(9), e183. http://doi.org/10.1371/journal.pmed.0050183

Pizzuto, J. (2002). Effects of dam removal on river form and process. *BioScience*, *52*(8), 683–691. http://doi.org/10.1641/0006-3568(2002)052[0683:EODROR]2.0.CO;2

Poff, L. N., & Hart, D. D. (2002). How dams vary and why it matters for the emerging science of dam removal. *BioScience*, *52*(8), 659–668. http://doi.org/10.1641/0006-3568(2002)052[0659:HDVAWI]2.0.CO;2

Ponder, W. F., Carter, G. A., Flemons, P., & Chapman, R. R. (2001). Evaluation of museum collection data for use in biodiversity assesment. *Conservation Biology*, *15*(3), 648–657. http://doi.org/10.1046/j.1523-1739.2001.015003648.x

Pringle, C. M. (1997). Exploring how disturbance is transmitted upstream: going against the flow. *Journal of the North American Benthological Society*, *16*(2), 425–438.

Puigbo, P., & Major, J. M. (2015). GPT: a web-server to map phylogenetic trees on a virtual globe. *PeerJ PrePrints*, *3:e1040*. http://doi.org/10.7287/peerj.preprints.840v1

PythonLabs. (2014). *Python*.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Sarkar, I. N. (2007). Biodiversity informatics: organizing and linking information across the spectrum of life. *Briefing in Bioinformatics*, *8*, 347–357. http://doi.org/10.1093/bib/bbm/037

Schauble, H., Marinoni, O., & Hinderer, M. (2008). A GIS-based method to calculate flow accumulation by considering dams and their specific operation time. *Computers and Geosciences*, *34*, 635–646. http://doi.org/10.1016/j.cageo.2007.05.023

Shaffer, H. B., Fisher, R. N., & Davidson, C. (1998). The role of natural history collections in documenting species declines. *Trends in Ecology and Evolution*, *13*, 27–30.

Shaw, S.-L., Yu, H., & Bombom, L. S. (2008). A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS*, *12*(4), 425–441. http://doi.org/10.1111/j.1467-9671.2008.01114.x

Sidlauskas, B., Bernard, C., Bloom, D., Bronaugh, W., Clementson, M., & Vari, R. P. (2011). Life in science. Ichthyologists hooked on Facebook. *Science*, *332*, 537.

Stamatakis, A. (2005). Phylogenetics: Applications, Software, and Challenges. *Cancer Genomics & Proteonomics*, *2*, 301–306.

Stanley, E. H., & Doyle, M. W. (2002). A geomorphic perspective on nutrient retention following dam removal. *BioScience*, *52*(8), 693–701. http://doi.org/10.1641/0006-3568(2002)052[0693:AGPONR]2.0.CO;2

Suarez, A. V., & Tsutsui, N. D. (2004). The value of museum collections for research and society. *BioScience*, *54*(1), 66–74. http://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

TCEQ. (2009). Dams and Reservoirs, Subchapter B: Design and Evaluation of Dams. In *Texas Commission on Environmental Quality*. Retrieved from http://www.tceq.state.tx.us/assets/public/legal/rules/rules/pdflib/299b.pdf

Texas Commission on Environmental Quality. (2013). *TCEQ Hydromaps*. Retrieved from http://www.tceq.state.tx.us/waterquality/tmdl/hydromaps.html

Texas Parks and Wildlife Department. (2012). *Texas Parks and Wildlife Department GIS Lab Data Downloads*. Retrieved from http://www.tpwd.state.tx.us/landwater/land/maps/gis/data_downloads

University of Texas at Austin. (2015). *Biodiversity Collections*. Austin, Texas, USA. Retrieved from https://integrativebio.utexas.edu/biodiversity collections/collections/ichthyology- fish

Verpoorter, C., Kutser, T., Seekell, D. A., & Tranvik, L. J. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters*, *41*, 6396–6402. http://doi.org/10.1002/2014GL060641

Vision, T. J. (2010). Open data and the social contract of scientific publishing. *BioScience*, *60*(5), 330–331. http://doi.org/10.1525/bio.2010.60.5.2

Vollmar, A., Macklin, J. A., & Ford, L. S. (2010). Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics*, *7*, 93–112.

Wang, L., Infante, D., Lyons, J., Stewart, J., & Cooper, A. (2011). Effects of dams in river networks on fish assemblages in non-impoundment sections of rivers in Michigan and Wisconsin, USA. *River Research and Applications*, *27*, 473–487. http://doi.org/10.1002/rra.1356

Whiteaker, T. (2014). *Fluviageny toolbox and documentation*. Austin, Texas, USA. Retrieved from http://tools.crwr.utexas.edu/Fluviageny/index.html

Whitelaw, E., & MacMullan, E. (2002). A framework for estimating the costs and benefits of dam removal. *BioScience*, *52*(8), 724–730. http://doi.org/10.1641/0006-3568(2002)052[0724:AFFETC]2.0.CO;2

Wilde, G. R., & Urbanczyk, A. C. (2013). Relationship between river fragment length and persistence of two imperiled Great Plains cyprinids. *Journal of Freshwater Ecology*, *28*(3), 445–451. http://doi.org/10.1080/02705060.2013.785984

Wiley, E.O., & Lieberman, B. S. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. Hoboken, NJ: John Wiley & Sons Inc.

Wu, F., Mueller, L. A., Crouzillat, D., Petiard, V., & Tanksley, S. D. (2006). Combining Bioinformatics and Phylogenetics to Identify Large Sets of Single-Copy Orthologous Genes (COSII) for Comparative, Evolutionary and Systematic Studies: A Test Case in the Euasterid Plant Clade. *Genetics*, *174*, 1407–1420.

http://doi.org/10.1534/genetics.106.062455

Yu, H., & Shaw, S.-L. (2004). Representing and Visualizing Travel Diary Data: A Spatio-temporal GIS Approach. Presented at the 2004 ESRI International User Conference, San Diego, CA: ESRI.

Zorn, T. G., Seelbach, P. W., & Wiley, M. J. (2002). Distributions of stream fishes and their relationship to stream size and hydrology in Michigan's lower peninsula. *Transactions of the American Fisheries Society*, *131*, 70–85. http://doi.org/10.1577/1548-8659(2002)1312.0.CO;2